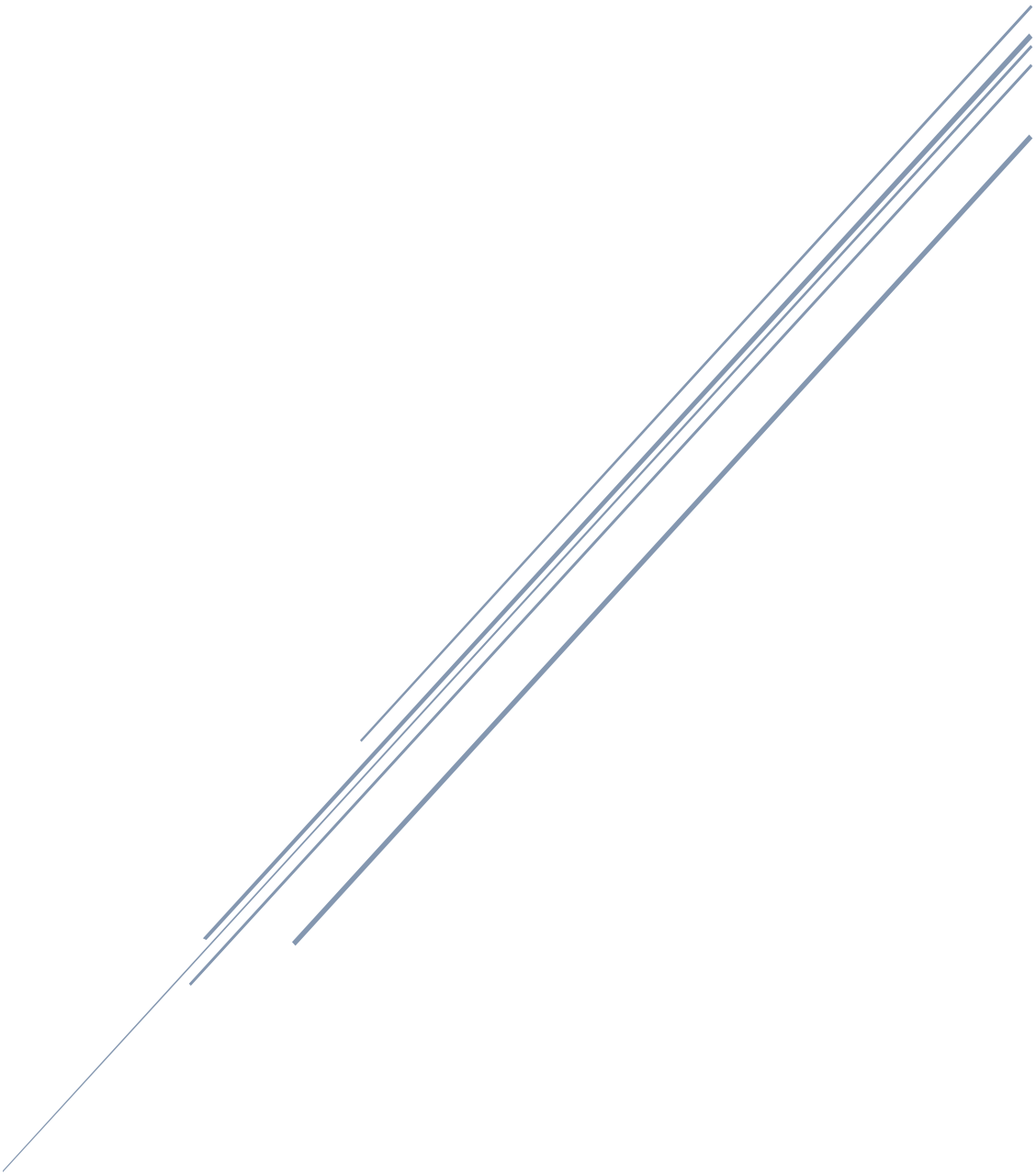
MOLECULAR BIOLOGY

Rodolfo Negri



Julian Politsch
La Sapienza

TABLE OF CONTENTS:

Syllabus:

1) The polymorphic DNA – Chemical components of DNA: bases, sugars and phosphodiesterase backbone (0.1 cfu). DNA B basic structure (0.2). Alternative DNA conformations, unusual structures (Cruciforms, triple helix) (0.2)

2) Conformational variability of structural parameters, curvature and bendability (0.2). DNA topology, winding and unwinding; linking number; DNA topoisomerases (0.2). The Genetic Code - the code decrypting; the structure and function of the code (0.1).

3) DNA replication: machinery and mechanisms in prokaryotes and eukaryotes (0.2). Replicons organization; topological and end-replication problems (0.1).

4) DNA mutability and repair; damage checkpoints (0.1).

5) DNA transcription: transcription in bacteria and bacteriophages; transcriptional machinery and RNA polymerase positioning signals (0.2). Methodological approach to the study of transcription: in vitro transcription systems (0.2). Transcription regulation in prokaryotic systems: activation and repressions; operon structures and function (0.2).

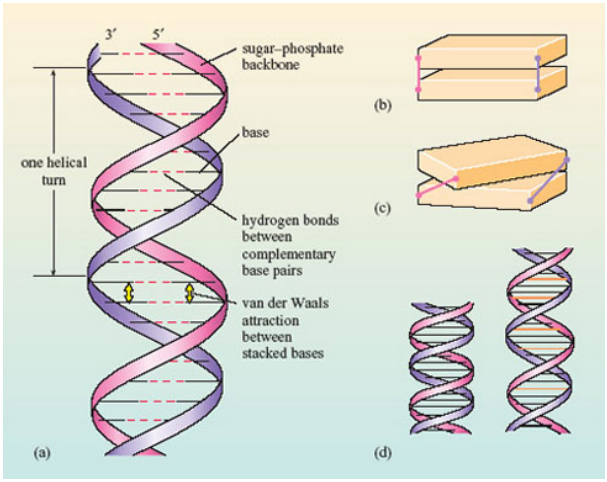
6) Transcription in eukaryotic systems (0.1); transcriptional machinery and RNA polymerase positioning signals (0.2). Methodological approaches to the study of transcription: eukaryotic in vitro transcription systems. Coordinate regulation of the three eukaryotic RNA polymerases (0.2). Transcription factors and transcription regulation in eukaryotes (0.2).

7) Molecular Biology Techniques (0.5 cfu): Nucleic acids purification, quantization, labelling and sequencing (introduction to NGS); PCR, RT-PCR, Southern and Northern blot; basic cloning techniques for analysis of protein-DNA interactions. Practical training (1 cfu).

8) Chromatin basic structure in eukaryotes and prokaryotes (0.2). The structure of nucleosomes and further organization levels (0.2). Histone modifications and their regulatory effects; the histone code. (0.2).

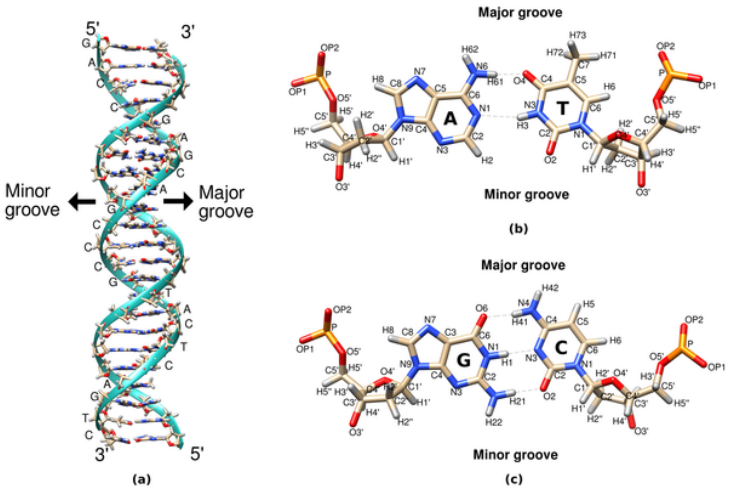
9) DNA methylation and its regulatory role (0.2). Histone variants and their regulatory role. Chromatin structure and chromatin remodelling at promoters (0.2). Chromosomes structure: centromeres, telomers and origin of replication (0.1).

Structure of DNA:



Double helix structure was quickly accepted due to its stability, replicability, and complementary base pairing. Base pairs are connected to each other by a rigid phosphate bond, which cannot be compressed. Between one base pair and another, water can enter, and therefore, when the stacked base pairs collapse they twist around each other. This hydrophobic interaction between consecutive base pairs, is called stacking, and stabilizes the base pairs, and forms the twist by solvating the water in-between them, through Van der Waal forces.

Another important feature of DNA is the existence of major and minor groves. Proteins can interact with major or minor grove by forming bonds with the bases, such as hydrogen bonds, electrostatic interactions and special covalent bonds. The existence of the major and minor groove is due to the antiparallel nature of the two DNA strands. Because of this glycosylic bonds are closer to one side of the double helix, and much further away on the major grove face. Major and minor grove alternate every 5 base pairs, and in most DNA it takes 10 base pairs to complete a full rotation. 80% of proteins interact with DNA in the major grove, due to the abundance of bonding elements, as well as the larger space.

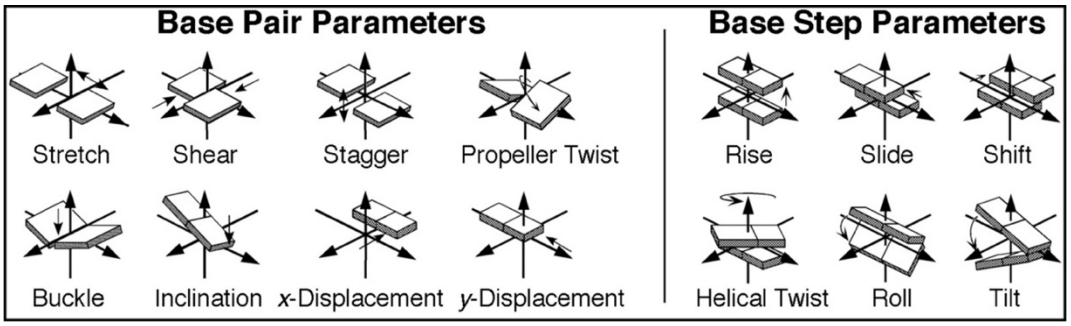


Base pairs are linked through hydrogen bonding between guanine and cytosine, and adenine and thymine. In G-C, three hydrogen bonds are formed, and in A-T, two are formed. Purines have two nitrogenous ring while pyrimidines have one and each base pair is made of a purine and a pyrimidine. You can remember purines with the mnemonic '**Pu**re **A**s **G**old' for adenine and guanine, and pyrimidines by thinking of a pyramid → sharp →**C.U.T.** for Cytosine, Uracil and Thymine. This model is seen in Watson and Crick, but another model -the Hoogsteen model- was proposed as an alternative way hydrogen bonds could form. The Hoogsteen model is preferred in acidic contrition, and therefor is not observed in cellular DNA, but can occur in triple helix elements, as well as RNA which is single stranded and single stranded DNA such as in telomeres.

Base Pair Parameters

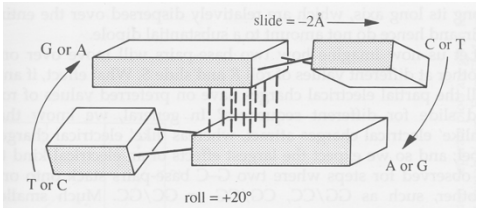
Base step parameters are the set of allowed motions observed in nucleic acid. Some movements are impossible such as rise, while others such as slide, twist and roll are allowed in small amounts. The amount of these movements necessarily must maintain the structure of the bonds, as we know phosphate bonds are rigid so rise, tilt, or shift which change the bond length are strictly not allowed. Together, they completely describe the relationship between the two blocks, in our case the two stacked base pairs.

Stacking occurs at a 36-degree angels, and also includes a propellor-twist between each base in a base pair, further decreasing the exposure to water, and increasing contact between two base pairs. The propellor twist is limited by the strength of the hydrogen bonds, twisting too much breaks the hydrogen bonds. AT pairs can twist more than GC because they only have 2 hydrogen bonds. The special case of repetitive A in one filament and T in the complement, there is an extremely strong propellor twist, and a 3rd hydrogen bond can form between the lower AT and the upper AT called a cross bond, or **bifurcated bond**, and this tight bonding blocks the condensation around the nucleosome, and keeping the chromatin open for transcriptional machinery.



Slide, the movement of the upper base pair along the bottom is possible, in small amounts, specifically 2 armstrong in either direction. Twist, that is rotations about the twist axis is easy and necessary -but depend in rotation- normally around 36-degrees but can differ greatly. Roll can be seen in variable amounts, but is limited by the strength of hydrogen bonds, and can be seen especially in chromatin packaging. These different base step parameters define the differences between A, B and Z DNA, discussed later.

The double helix shape is formed because the base pairs are highly hydrophobic, and therefor want to pile together to remove the water. But the base pairs are linked by a phosphate bond, which is resistant to compression. To overcome this, the DNA molecule base pairs twist about 36-degrees to maintain the length of the phosphate bond, while closing the space. The four nitrogenous bases are attached to the deoxyribose molecule by a glycosylic bond. Since the DNA strands are anti-parallel, and the glycosylic bonds vary in distance, we get a major and minor groove.



Parameter Manipulation: School Cases

1) CA/TG: In the case that a pyrimidine is followed by a purine, such as T followed by A, we always see positive slide involved in order to avoid steric clash, to compensate for the difference in size of a purine and a pyrimidine. It can also work if the slide is in the opposite direction, as seen in the picture, the steric hinderance will be avoided due to the increase in overlapping of the purines. ±20 Roll
2) GG/CC: Slide can also solve the problem of the partial charge on the base pair. While the base pairs are overall neutral, the difference in charge due to the distribution of H, C, N, O. This problem is mainly when a G is followed by another G This can be solved in the same manner, a positive or negative slide to align the opposite charges instead of creating repulsion. +0 to +10 roll
3) AA/TT: In this case, the bifurcated bonds make the AA/TT extremely rigid, therefor we have 0 slide and 0 roll.

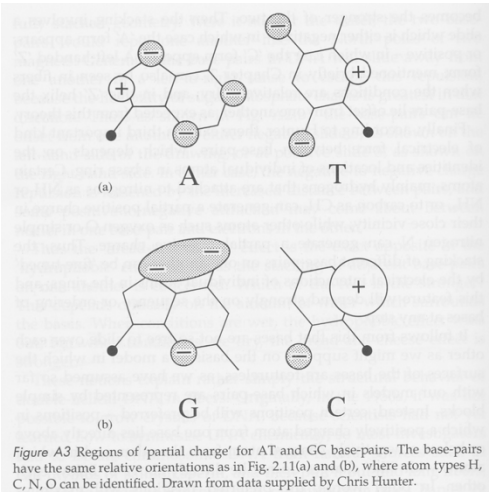
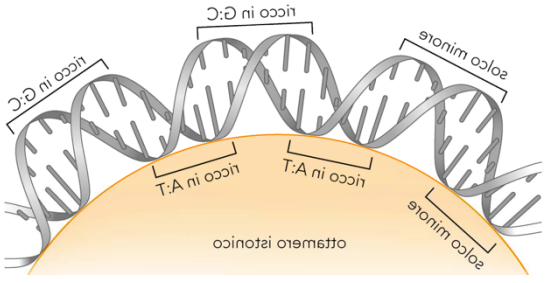
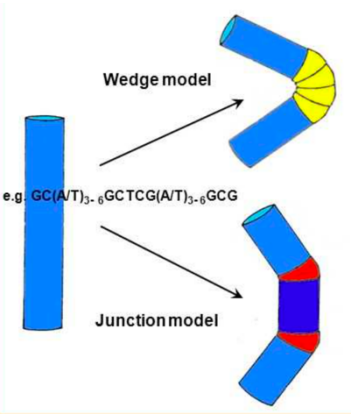


Figure A3 Regions of 'partial charge' for AT and GC base-pairs. The base-pairs have the same relative orientations as in Fig. 2.11(a) and (b), where atom types H, C, N, O can be identified. Drawn from data supplied by Chris Hunter.

Positive roll opens the minor groove, negative roll opens the major groove. Therefor we can see if roll alternates between positive and negative at a variable rate, we can get the same curvature of the DNA. A completely smooth curve follows a sin graph pattern. The sequences to be opened also need to be in the correct place, as we know roll is impossible in TT/AA sequences. Cells contain intrinsically curved DNA, which also contains a difference in mobility in gel electrophoresis due to the inability to move through the channels in the gel. Therefore, DNA molecules of the same length, only differing in the intrinsic curvature, move at different speeds though a median.



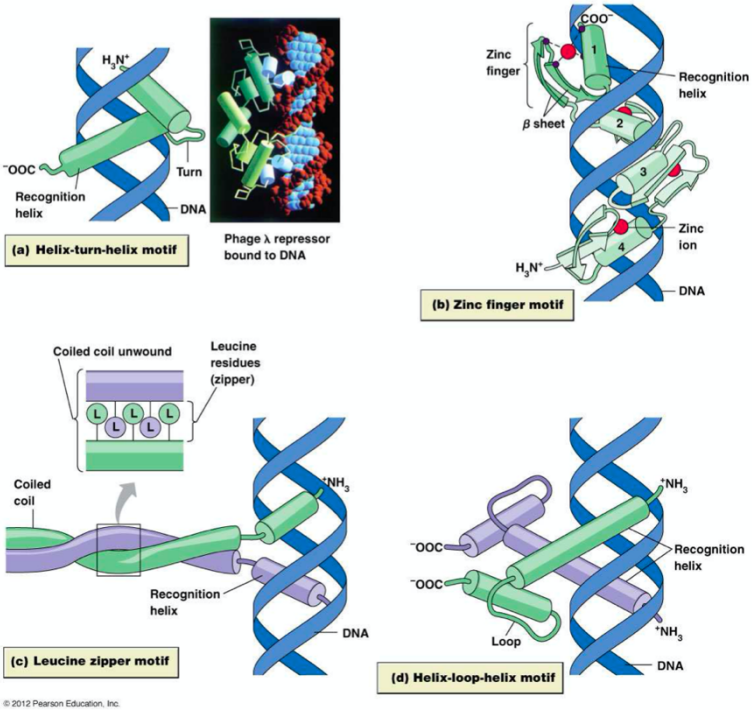
Models proposed for the intrinsic curvature of DNA are junction, and wedge model. In junction model, it was hypothesized that a-tracks were found in every 6/10 DNA molecules, which contain AA/TT base pairs thus blocking curvature in specific sequences, therefor a consistent, repetitive pattern of bendable and unbendable serves to permanently curve the DNA. In bacteria, the coding sequences are usually very straight(AA/TT) rich, and the promoter has as high tendency to curve, due to the need to wrap around the polymerase molecules.



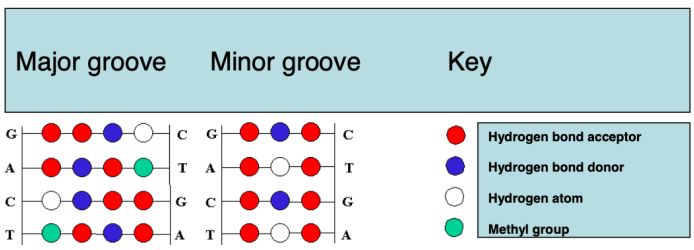
Curvature can also be created by proteins, as long as the DNA is flexible. This is best seen in the nucleosomes, in which the histone proteins in the disc interact in the minor groove to wrap the DNA. Since the interaction is always with the minor groove, the major groove is enlarged and exposed, creating negative roll. Therefor nucleosome DNA should prefer AA/TT in specific points, away from the curvature, and CG in the places where curvature is very strong. This was tested through the purification of nucleosome DNA through cocci digestive enzymes, which cannot digest nucleosome DNA but can with free DNA. In nucleosomes, therefor, rigid(AA/TT) and flexible(ATAT) base pairs alternate every 5 base pairs.

DNA binding motifs

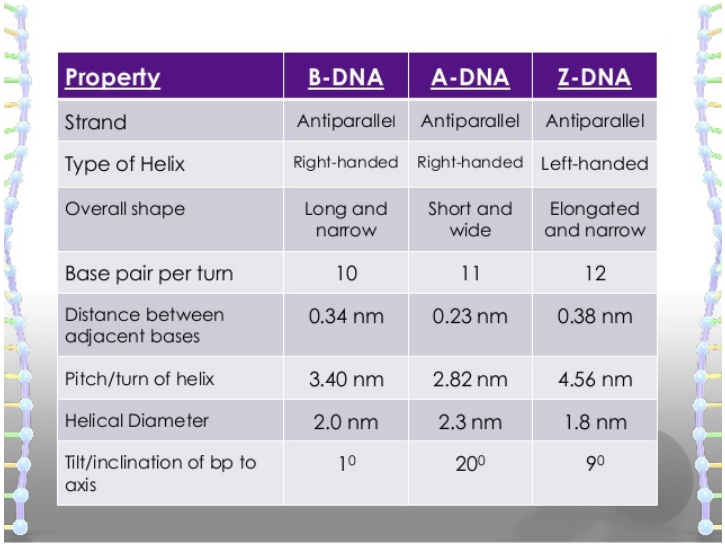
Protein alpha helixes enter the DNA groves to bind the bases, in which the recognition helix is always inserted into the major groove. Base pairs offer different chemical groups that allow the proper recognition helix to be tightly associated with the groove, through hydrogen bonds, methyl bonds(hydrophobic interactions), or hydrogen atoms. The tendency to bond with the major groove is because the major groove contains 4 possible bonding elements, while in the minor there are only 3, which are summarized in the table below. The variety of bonding elements in the major groove allows for more specific interactions, as well as a larger variety of possible bonds to be formed. The recognition helix is the general binding factor, but the specific, tight interactions come from the larger DNA binding motif.



Some DNA binding proteins use **cooperative binding** to increase their stability while binding to DNA, either though binding to another copy of the same protein, or another. This can be tested thought tracking of the speed of movement of DNA through gel, decreasing when the protein is bound. This test is used to see if the speed is linearly related to protein concentration, in the case that it isn't, we likely have cooperative binding occurring.



Unusual forms of DNA



Property	B-DNA	A-DNA	Z-DNA
Strand	Antiparallel	Antiparallel	Antiparallel
Type of Helix	Right-handed	Right-handed	Left-handed
Overall shape	Long and narrow	Short and wide	Elongated and narrow
Base pair per turn	10	11	12
Distance between adjacent bases	0.34 nm	0.23 nm	0.38 nm
Pitch/turn of helix	3.40 nm	2.82 nm	4.56 nm
Helical Diameter	2.0 nm	2.3 nm	1.8 nm
Tilt/inclination of bp to axis	1^0	20^0	9^0

DNA can be seen in three forms, B, A, and Z. B is the 'standard' form we find in most biological processes. In B, there is minimal slide and tilt. The A form is typical in the absence of water, down to 60%. In A form, the shape is completely different, firstly, the diameter is larger than the B-DNA, and the same number of base pairs is shorter in A. Secondly, the major and minor groove look approximately the same in the A confirmation. Lastly, we can see that the bases are not perpendicular to the axis and are instead inclined about 18degrees. This causes the bases to not pair up in the center of the molecule. The third confirmation, Z or alternative confirmation, is observed in the perfect repetition of GC/GC and is completely different from B and A DNA. In Vivo, the B confirmation can form when RNA and DNA pair, in the lab. Z can form in the presence of high ionic strength, negative supercoiling, and cytosine metalation.

The most puzzling difference Is the left-handedness of Z-DNA. It is explained by the conformation of the glycosylic bond, which can exist as either anti, or syn. To go from one to the other, we have to flip the bases by 180 degrees. In A and B, the confirmation of all bases is always anti. In Z however, the G is always in syn, while the C is always in anti, so the location of the glycosylic bond slips back and forth every other base, creating a zig-zag shape of the phosphodiester filament. This zig-zag leaves holes in the side of the molecule, allowing water to enter, and thus methylation of the cytosines closes the holes and stabilizes the z-DNA

Z-DNA can be formed In vivo in very special cases, such as when there is huge negative supercoiling, which favors the lefthanded formation of DNA in GCGC sequences.

Another unusual structure that can occur in DNA is the triple stranded helix, where two strands are forming Watson and Crick base pairs, and the third is forming Hoogsteen base pairs. Triple helixes can be found inside the cell in some replication intermediates but can also be studied in the lab to study transcription factors, or other DNA binding elements, in order to design elements to form triple helixes inside the cells.

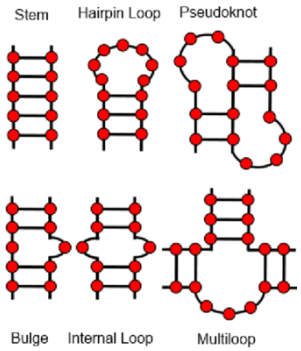
Another important structure in eukaryotic DNA is the single stranded telomere, which needs to be protected. The end of the telomere contains repeating Guanines, which can form a quartette, formed by Hoogsteen base pairing, eventually forming into baskets, that protect the single stranded telomeric DNA.

We can also study the hairpin structure form in palindromic DNA, separated by a non-repeating sequence, thus the DNA can reassociate to itself instead of the complementary strand. The cruciform can form when the stem is very long. The loops must be small enough to make the energy cost possible, but long enough to loop around itself. Cruciforms can be created again by negative supercoiling. RNA is very likely to form cruciform, because they are actually the more stable confirmation for single stranded RNA.

RNA-Structure

RNA structure is much freer to form strange structures, because it's a single stranded molecule and bends, rotates and binds with itself freely. The primary structure of RNA is defined by the nucleic acid sequence, G C U A These folding shapes, such as stems, hairpin loops are examples of RNA secondary structure. Tertiary

structure is the association of the DNA secondary structures around itself, as well as the quaternary structure formed by different RNA associating with each other, such as in ribosomal RNA.

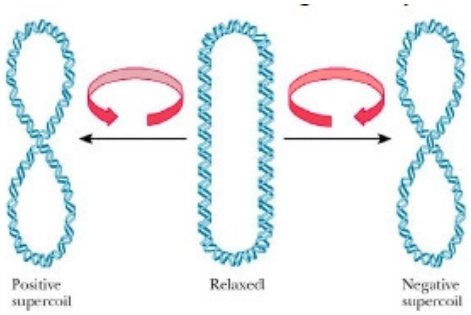


The three-dimensional shape of RNA can be predicted through 1) the computational evaluation of the lowest energy confirmation, and 2) the known motifs contained in the RNA that can be experimentally determined. Most predicting software are a mix of these methods.

DNA topology and supercoiling:

Supercoiling refers to the twisting of the DNA double helix, can be negative or positive. Topology is the study of systems that do not change when they are deformed, think topographic maps. Therefore, topologic systems are closed systems, that resists deformation. DNA is a topological system, because in presence of deformation, it spreads this deformation out. Single stranded linear DNA has topological domains, which are defined by Tw (Twisting number) referring to helical turns, Wr (whirring number) and Lk (linking number) as well as ó (supercoiling) N is number of nucleotides. The law of topology tells us that linking number is constant, so increasing twist must result in decreased wiring, vice versa, in a closed system. Supercoiling is the systems response to constant twisting, and results in supercoiling in the absence of opposite twists. This occurs only when both ends of the domain are fixed, such as in circular DNA or in linear DNA where the extremities cannot rotate, called a closed system. Negative supercoiling, is leftwards, and is created by rightward twisting of the DNA, while twisting leftward creates positive super coiling.

$$Lk = Tw + Wr$$
$$\Delta Lk = \Delta Tw + \Delta Wr$$
$$Tw = N/10.5$$
$$\sigma = \Delta LK/N$$



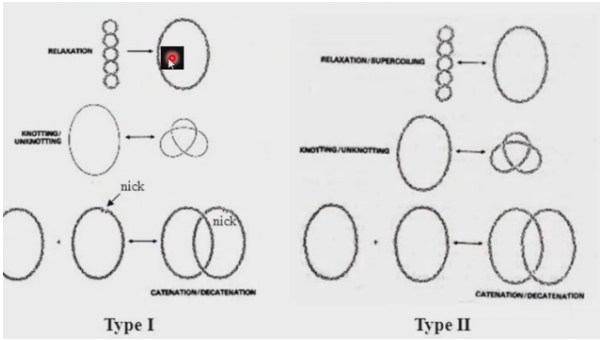
Bacterial DNA and plasmids are constantly in a twisted, supercoiled form, and are invariably, always, negative. This is understood through the formula, if we separate the DNA strands in bacteria, we see the reaction in the relaxed DNA exactly opposite and equal, for example the loss of 40 nucleotide pairs creates 4 elements of positive supercoiling. In a supercoiled loop of DNA, the opening of the DNA releases the supercoiling, an energetically favorable reaction. DNA in negatively supercoils serves as a storage for free energy, and makes the DNA strands easier to separate, that's why circular DNA is always negatively supercoiled, except in the few species who want to prevent denaturation(high heat). Supercoils can be *plectonemic*, where the twist is right-handed, and *toroidal*, which is found in eukaryotes, where the supercoils are wrapped around something, mainly histones.

Topoisomerases

DNA Topoisomerases modify the topology of systems, by breaking a closed system, then resealing this open system. Topoisomerases are broken into:

- Topo I: breaks only one of the filaments
- Topo II: breaks both DNA filaments

Topo I relax the system when one strand is already contains a nick, because it can only cut one strand, while Topo II can relax or increase supercoiling by itself. DNA-Gyrases in bacteria can increase supercoiling from relaxed DNA using the energy of ATP. Topo I and Topo II can also add or remove knots in DNA, as well as catenation and decatenation. Catenation and decatenation is especially important in bacterial reproduction.



Electrophoresis allows us to distinguish and identify the different topological forms, called topoisomerase. This relies on the same principle as naturally curved DNA, where supercoiling elements move faster through the channels of the median compared to relaxed circular DNA. It was also experimentally determined that topoisomerase with increasing number of supercoils increases movement speed through a median. Supercoiling can be created in the lab by introducing DNA binding elements such as ethidium bromide, which decreases the twist, but will be relieved by topoisomerase. The removal of the topoisomerase and the treatment agent will close the space between base pairs, and the decreasing twist will promote the positive the supercoiling.

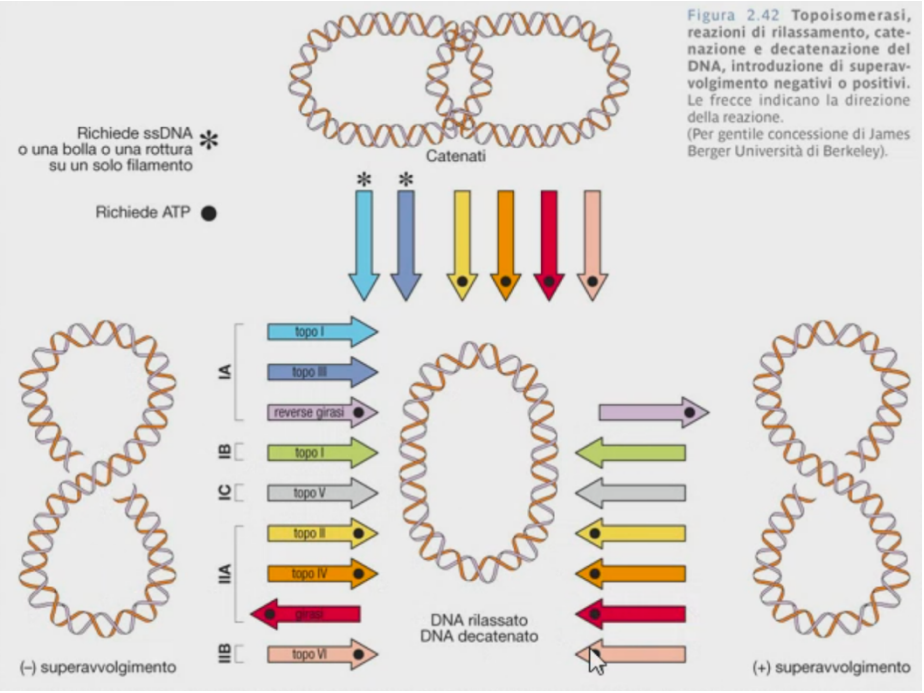


Figura 2.42 **Topoisomerasi, reazioni di rilassamento, catenazione e decatenazione del DNA, introduzione di superavvolgimento negativi o positivi.** Le frecce indicano la direzione della reazione.
(Per gentile concessione di James Berger Università di Berkeley).

Topoisomerases break the dystric bond while simultaneously forming another diester bond to an attached tyrosine, forming a phosphotyrosine intermediate (1 in topo I and 2 in topo II) which will later be used to reclose the DNA and reform the diester bond. In this way they don't lose energy when breaking and reforming the bonds.

Topo I can be further classified into free rotation, seen in eukaryotes, and bridge in bacterial. Free rotation works by the breaking of one filament in supercoiled DNA, and the release of the supercoiling around the freely rotation single strand. This kind of mechanism can relax positive, or negative supercoils, as the relaxed form is lower energy than supercoiled. In the bridge by contrast, the process is active. First, one strand is broken, then the DNA is crossed, in order to allow rotation only in one direction. This specialization allows the bacteria to only relax negative supercoiling, instead of positive and negative.

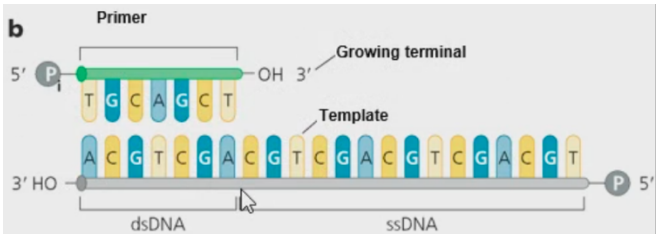
Topo I and III are found in bacteria and relax only negative supercoiling. Reverse gyrase, in thermophile prokaryotes releases negative supercoiling AND increases positive.

The bridging mechanism in topo I is well studied. It begins with the transfer of the dystric bond to a tyrosine, followed by a strand passage across the protein, which is then rejoined. Topo II instead uses the gate model, which works in a similar way, with the double breakage, the fixation of one strand, followed by the relegation.

Twin-supercoiled-domain model

James wang proposed the *twin-supercoiled-domain model*, that states what when a complex moved along DNA, it will create positive supercoils on the front, and negative supercoils on the back. This is especially important in RNA polymerase, as it moves towards one direction, it creates a region of positive supercoiling in the direction of transcription, and an underwound region of negative supercoiling behind it. Even though the linking number remains the same, the distribution of supercoiling is dense in some domains. Due to the accumulation of super helical turns, the process of transcription is made difficult. For this reason, topoisomerases need to constantly work to relax the DNA in different domains.
The transfer of the phosphate to topo-I, to form the phosphotyrosine intermediate is a reversable process, and therefor conserves the bond energy from breaking and reforming the bond. In cancer research, one major field of study is to target the topoisomerases, so they cannot reseal the breakage. Chemotherapy drugs target the cleavable complex for both topo-II and topo-I, specifically, blocking the cleavage of the DNA, stopping transcription, inducing intrinsic p-53-caspace-9 apoptosis.

DNA Replication:



Basic concepts of replication

DNA replication is known to be semi-conservative. Replication can only occur from 5' to 3', and always requires a primer. This means that DNA polymerases need a primer nucleotide, specifically with a 3' -OH growing terminal. The need for the primer is also understood by the presence of a 3rd phosphate at the 5' end, so we can only add nucleotides to the 3' end. If this was not the case, energy would be lost in order to reform the 2nd-3rd phosphate bond with every addition. In the case of adding to the 3' end, an enzyme called pyrophosphatase hydrolyzes the triphosphate, exposing one phosphate which can be added to the -OH 3' end easily.

The complex process of DNA replication is mediated by enzyme complexes called DNA-polymerase, in the case of prokaryotes DNA polymerase-III. DNA polymerase can be described as a hand, where the incoming chain enters through the palm, and the fingers and thumb wrap around. The complexity ensures that the wrong nucleotide is extremely difficult to add, where the incoming nucleotides must be exactly the right shape and polarity in order to not make steric hinderance. Additionally, the wrong nucleotide will result in the phosphate bond being in the wrong place, and thus will not form. There is also a control for DNA vs RNA, in RNA, the -OH groups position will create steric hinderance and thus will be unstable. DNA polymerases also require divalent cations to function, which regulate the acceptance or rejection of an added nucleotide. Hydrogen bonds also insure the correct base pairing. In the case that the dNTP is a good fit both in the metal ion pocket, and in formation of complementary hydrogen bonds, the O-helix of DNA Pol will clamp down like a hand gripping and simultaneously cleave two phosphates, while binding the remaining phosphate to the free 3' OH group.

DNA polymerization is a fast process, because it is a processive, meaning once a polymerase is attached, it adds many nucleotides consecutively without disassociating. It is also an extremely accurate process because of the three principles of:

1) DNA polymerase selectivity (1/million mutation)
2) Proof reading activity (1/10 million mutation) by 3' exonuclease activity, removing the base pair that was just added at a much higher rate when it is a mismatch
3) Error repair (1/1 billion, final frequency of mutations in humans)

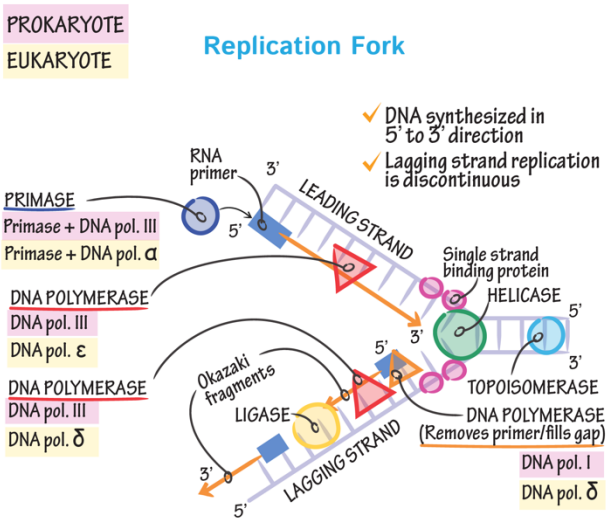
Error repair involves first the detection of a mismatched base pair. After identification, the DNA is transferred from the catalytic site, to the repair site, where the wrongly paired DNA is removed, and fixed, and is them transferred back.

DNA replication enzyme

In DNA replication, there is a leading strand, and a lagging strand. In the leading strand, replication is continuous, while in the lagging it is discontinues, where instead multiple primers are needed to attach every new Okazaki fragment. Many activities are needed in replication, it's very important to remember these (Test question):

We need many different activities for efficient replication.

1. A **helicase** activity to break the non-covalent bond between base pairs
2. **Single strand binding** proteins to keep the separated DNA from rejoining or folding
3. **Initiator protein** to recognize origin of replication (DnaA in Pro, ORC in Euk)
4. **DNA polymerase** to synthesize new strands



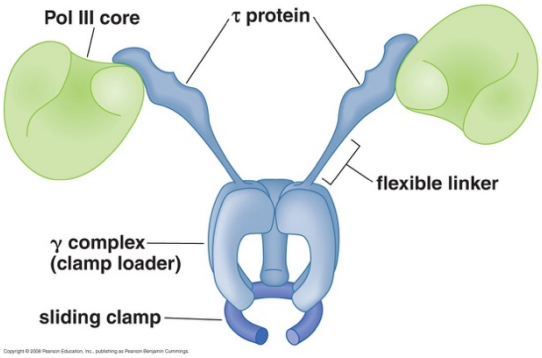
5. **DNA ligase/polymerase** to reseal the lagging strand okazaki fragments, and remove primers
6. **Clamp protein** to keep the polymerase secured and oriented
7. **Clamp loader protein** to load the clamp around the DNA
8. **RNA primase** activity to start new DNA strands from
9. **Topoisomerase** Activity to relieve topological strains (according to the twin supercoiled domain model)

The factors of origin recognition vary between prokaryotes and eukaryotes. In prokaryotes, the binding of the initiator protein induces the opening of the AT rich DNA sequences and the initiator is a single protein. In Eukaryotes, it is many proteins forming the ORC.

The helicase runs ahead of the replication fork, and when the polymerase advancement is slower than the helicase advancement, helicase activity is reduced. In this way the helicase is regulated by the speed of the polymerase activity, and does not unnecessarily unzip DNA too far ahead.

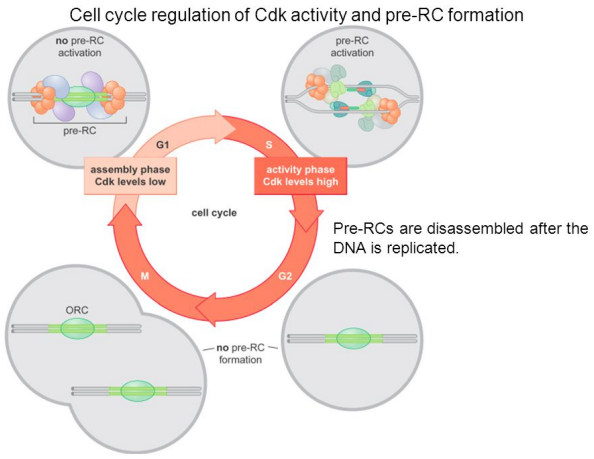
In the leading strand there is only one polymerase and one sliding clamp involved, while in the lagging strand the polymerase synthesizes an okazaki segment and disassociates, while the sliding clamp remains attached as a bookmark of sorts. Sliding clamps are conserved through evolution, where the ring structure, and the diameter is the exact same, even when the proteins are different. This suggests the need for such a mechanism no matter the size of the genome.

At the replication fork, a clamp loader protein attached to two linker proteins connects the two DNA pol which will synthesis the leading, and lagging strand. This complex of clamp loader, polymerases, linkers, and clamps is called the pol holoenzyme. In this way, the helicase activity runs in front of the holoenzyme and feeds the ssDNA templates to both the leading and lagging strands, while the clamp loader can load for either strands. The combination of all the proteins that function at the replication fork is called the replisome.



Regulation of Replication

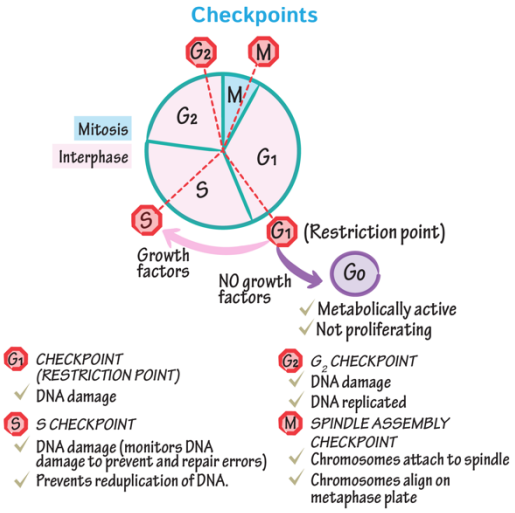
It is important to regulate replication in order to duplicate DNA before entry into the M-Phase. In prokaryotes, the timing of replication relies on methylation of A (C is methylated in Eukaryotes, but for transcription repression not replication regulation). The fully methylated sequences, produced by Dam-methylase -in the GATC islands- is the main signal for replication initiation. In this process, multiple initiator proteins called DnaA binds to the replicator sequence (origin of replication in eukaryotes), and promote strand separation with their SSB activity. Due to the semi conservative nature of DNA replication, the two new strands will be formed of one methylated strand, and one new, unmethylated stand. These hemi-methylated chromosomes will attract a protein called SeqA, which upon binding, blocks replication and ensures the region will not be replicated again. After some time, SeqA disassociates and DAM protein is allowed to methylate the new strand, returning both sister chromosomes to their fully methylated state. In the final stages of replication in prokaryotes, topo-II induces a two stranded break, disconnecting the two chromosomes.



In eukaryotes, methylation does not play this role. Instead, cyclin dependent protein kinases are cyclically activated and deactivated/destroyed to ensure replication happens exactly once. Replication selection occurs exclusively in

G1 because origin recognition complex (ORC) assembly happens in this phase, due to specific kinase activity recruiting Cdc6 and Cdt1. In this phase, the bound factors act as repressors for replication firing and anchor the complex at the origin of replication. Next, in S phase replication is fired at the ORC (now called Pre-RC) due to S-CDKs ability to phosphorylate and remove Cdc6 and Cdt1, allowing the complex to fire and replication to begin on various points in the genome.

Cell Cycle and control



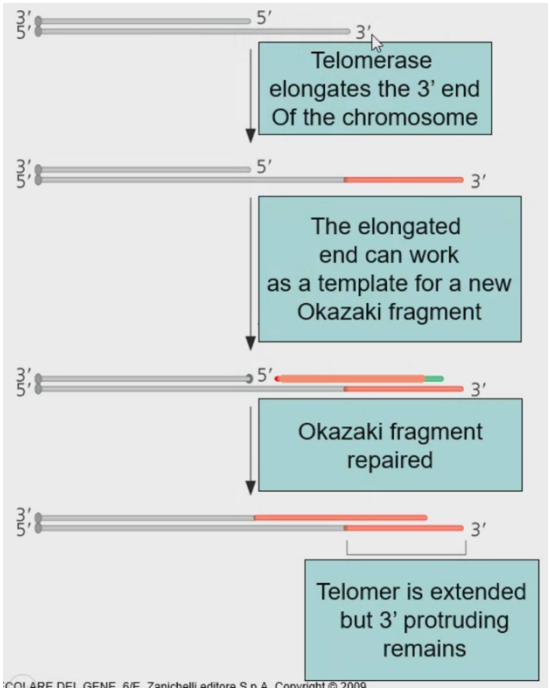
The cell cycle is regulated by CDKs. Several cyclins peak in concentration and activity at different points in the cell cycle. The cell cycle contains many checkpoints, which is a window of time when a cell must decide if it can proceed into the next stage or change course. In the G1 checkpoint, cells must be large enough, be receiving growth factors, as well as proper nutritional state in the cell. At G1 there is also a DNA damage control check point, if DNA is every damaged, the cell won't move to the next phase. In the G2/M checkpoint, the cell makes sure replication is complete, as well as that there is no DNA damage. The last checkpoint, the spindle checkpoint occurring in the M-phase, and is regulated by the APC/C, which can halt replication in the case that the chromosomes are not correctly attached at the metaphase plate.

Cyclins and CDKs are the control systems of the cell, where the activation of the CDKs depends on the concentration of cyclin, which is controlled by transcription. We also have CDK inhibitors such as P16, P12, which block the interaction of cyclin with CDKs. Many other control systems such as Wee1 (kinase) and CDC25 (phosphatase) which activate or deactivate cyclins.

DNA damage checkpoints are essential for the maintenance and safety of the genome. DNA damage is sensed by special protein sensors, who activate transducers (kinases), who activate effectors, who create a specific response. Effectors immediately stop the cycle, then decide if the cell can repair the damage , or enter apoptosis. Alternatively, the DNA damage signal tells the cell to transcribe proteins that work in the repair system and stop the cell cycle. The cell cycle control system is massively complex, models are nearly impossible to create accurately because there are too many moving enzymes and pathways to accurately map. In theory we can make a model of the cell cycle control systems, but many of the components are unknown, as well as cell variability ruins the usefulness of any one model. The system can be easily modeled by the understanding of simple circuits, where an input A can wither activate or deactivate B. This concept is called a simple oscillator, and it mirrors how CDK activity is regulated in an oversimplified way.

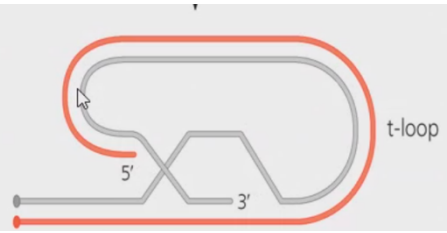
Telomeres, centromeres and origin of replication

The cell encounters a problem at the end of the chromosome when replicating the last sequences, as they cannot add a primer to the 5' extremity, a problem that gets worse and worse with every replication. This would mean the cell loses genetic information every cell cycle. This problem is solved in some viral cases by a protein that can prime the end of the chromosome.



In Eukaryotes however, a protein complex called a telomerase, which contains a catalytic RNA, makes a template, that can replicate the extremity of the chromosome, by repeating the same sequence over and over and then adding identical segments. This

repair system still ends with a protruding end on the 3' strand, which is highly likely to be damaged, especially because it is repetitive it can easily be recognized as the wrong sequence. The cell must protect these extremities, forming a region called the telomere. Sheltered proteins bind to the telomeres. These proteins differ in different animals. In yeast, these sheltering proteins regulate the activity of the telomerase enzyme, more copies create more inhibition of the telomerase. In mammals, it is theorized that the single stranded extremity would base pair to a piece of the double stranded sequence, forming a tetraloop (t-loop), further protecting the single strand. Since the telomere sequences are rich in guanine, they can also likely form g-baskets.



Telomeres are constant in length in germline cells, somewhat stable are stem cells, while somatic cells see a rapid shortening of their telomeres. Cancer instead can restabilize the telomerase activity, allowing them to replicate indefinitely. It seems many other diseases are related to telomere length; longer telomeres show higher resistance.

Centromeres are the genomic site in which the kinetochore is formed, and thus the place where the microtubules attach in mitosis. Centromeres are usually very AT rich regions who bind proteins, and special nucleosomes. These nucleosomes are important for the contact and connection of the spindle fibers, while the proteins assist in attraction and docking as well. In yeast, we have studied the origin of replication through splicing and experimental transcription.

DNA Repair:

Mutations

Replication errors

Replication errors can be transitions, where a pyrimidine to pyrimidine or purine to purine shift occurs (T to C, A to G) or transversions, which a pyrimidine-to-purine occurs or vice versa (T to G, A to C). Other mutations that insert or remove a single (or small number) of nucleotides also can occur; these mutations are all together called point mutations. Point mutations can be; frameshift, missense (wrong amino acid), nonsense (premature stop) or silent (Base changed, but same amino acid).

A slippage of a triplet can lead to a triplet being excluded or added. Nucleotide deletion and insertion resulting in frameshift is very deleterious, because they necessarily affect all the codons downstream, leading to missense and eventually nonsense codons. A deletion followed by an insertion is less likely to be a deleterious mutation but can still affect the polypeptide especially if the insertion and deletion are far apart. Some regions of the chromosome are 'hotspots' for slippage which usually contain large amounts of short repeats called microsatellites.

DNA Damage

Exogenous DNA damage can come from many sources:

1) UV Light can cause:
    a. Deamination:
        i. Of cytosine forming the base Uracil
        ii. Of 5-methyl-cytisine forming thymine
    b. Depurination: the spontaneous hydrolysis of the N-glycosyl linkage, forming an apurinic base (no sugar)
    c. Formation of crosslinks In neighbor T bases, forming thymine dimers
    d. Formation of reactive oxygen agents, able to oxidize Guanine to 8-oxoG. oxoG is highly mutagenic because it can base pair with A and C.
2) X-ray/Gamma can cause double strand breaks:

    a. Ionization radiation from x-rays can directly attack the deoxyribose in the backbone
    b. Or it can cause damage indirectly by forming oxygen free radicals, indirectly contributing to instability
3) Mutagenic agents:
    a. Base analogues are chemically similar to bases, but do not perform the same functions (5-bromouricil for cytosine)
    b. Intercalating agents which are able to slide in-between the base pairs (Proflavine)
4) Alkylating chemicals such as nitrosamines alkylate (usually methylate) the oxygen attached to C6, creating O6-methylguanine, which often mis-pairs with T.

Normally suppressor tRNA recognizes the stop codon, but a mutated tyrosine-tRNA recognizes the wrong element and terminates prematurely. One case where this does not happen is when the tyrosine-tRNA is mutated in the same way as the codon, allowing a selective advantage.

SS Damage

Direct Reversal

Cells are known to eliminate three types of damage to their DNA by chemically reversing it. These mechanisms do not require a template, since the types of damage they counteract can occur in only one of the four bases. Such direct reversal mechanisms are specific to the type of damage incurred and do not involve breakage of the phosphodiester backbone. The formation of pyrimidine dimers upon irradiation with UV light results in an abnormal covalent bond between adjacent pyrimidine bases. The photoreactivation process directly reverses this damage by the action of the enzyme photolyase, whose activation is obligately dependent on energy absorbed from blue/UV light to promote catalysis. Photolyase, an old enzyme present in bacteria, fungi, and most animals no longer functions in humans, who instead use nucleotide excision repair to repair damage from UV irradiation. Another type of damage, methylation of guanine bases, is directly reversed by the protein methyl guanine methyl transferase (MGMT), the bacterial equivalent of which is called ogt. This is an expensive process because each MGMT molecule can be used only once; that is, the reaction is stoichiometric rather than catalytic.[18] A generalized response to methylating agents in bacteria is known as the adaptive response and confers a level of resistance to alkylating agents upon sustained exposure by upregulation of alkylation repair enzymes.[19] The third type of DNA damage reversed by cells is certain methylation of the bases cytosine and adenine.

Mismatch repair

Mismatch repair systems are present in essentially all cells to correct errors that are not corrected by proofreading. These systems consist of at least two proteins. One detects the mismatch, and the other recruits an endonuclease that cleaves the newly synthesized DNA strand close to the region of damage.

In bacteria, MutS senses the mismatch because the mismatch will distort the structure of the double helix. MutS then recruits MutL and MutH. MutH cuts the mismatched pair, and an exonuclease digests the mismatched pair, while the DNAPol-III replaces it. This process can work from 5’ to 3’ or the other way. The prokaryotic DNA knows to replace the old strand because the old strand is methylated, therefor the nick and repair will only happen on the old strand. The similar process happens in Eukaryotes, with different enzyme name.

Base Excision repair (BER)

damaged single bases or nucleotides are most commonly repaired by removing the base or the nucleotide involved and then inserting the correct base or nucleotide. In base excision repair, a glycosylase enzyme removes the damaged base from the DNA by cleaving the bond between the base and the deoxyribose. These

enzymes remove a single base to create an apurinic or apyrimidinic site (AP site). Enzymes called AP endonucleases nick the damaged DNA backbone at the AP site. DNA polymerase then removes the damaged region using its 5' to 3' exonuclease activity and correctly synthesizes the new strand using the complementary strand as a template. The gap is then sealed by enzyme DNA ligase.

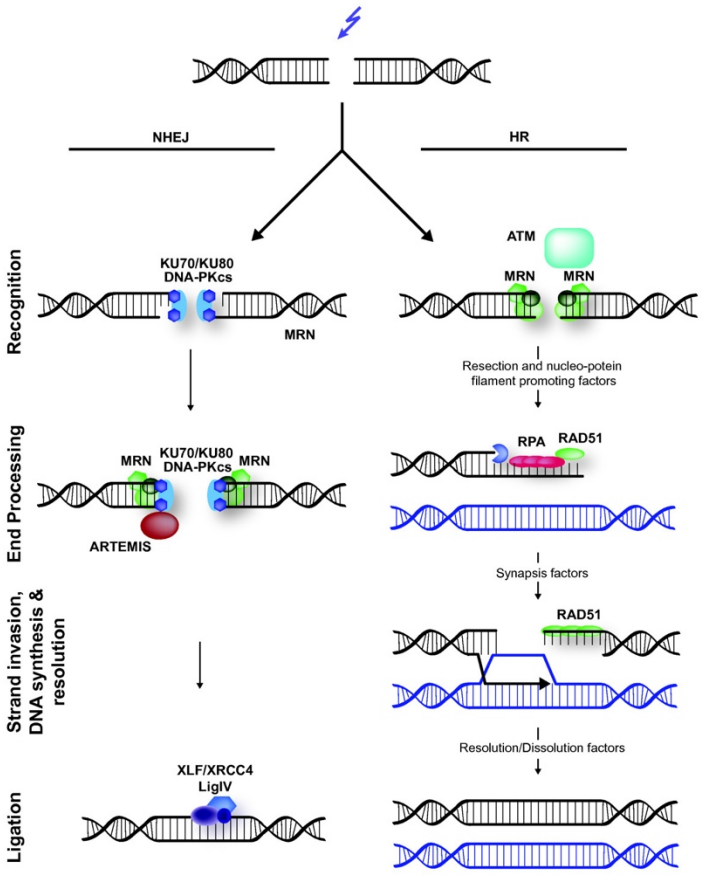
Nucleotide Excision Repair (NER)

Bulky, helix-distorting damage, such as pyrimidine dimerization caused by UV light is usually repaired by a three-step process. First the damage is recognized, then 12-24 nucleotide-long strands of DNA are removed both upstream and downstream of the damage site by endonucleases, and the removed DNA region is then resynthesized. NER is a highly evolutionarily conserved repair mechanism and is used in nearly all eukaryotic and prokaryotic cells. In prokaryotes, NER is mediated by Uvr proteins. In eukaryotes, many more proteins are involved, although the general strategy is the same. In bacteria, excision repair is called UV SRS repair but works on the same principle of incision and removal.

DS Damage

Homology directed repair (HDR)

HDR is a mechanism in cells to repair double-strand DNA lesions. The most common form of HDR is homologous recombination. The HDR mechanism can only be used by the cell when there is a homologous piece of DNA present in the nucleus, mostly in G2 and S phase of the cell cycle.



NHEJ

Non-homologous end joining (NHEJ) is a pathway that repairs double-strand breaks in DNA. NHEJ is referred to as "non-homologous" because the break ends are directly ligated without the need for a homologous template, in contrast to homology directed repair, which requires a homologous sequence to guide repair. NHEJ is especially important before the cell has replicated its DNA, since there is no template available for repair by homologous recombination. NHEJ is typically guided by short homologous DNA sequences called microhomologies. These microhomologies are often present in single-stranded overhangs on the ends of double-strand breaks. When the overhangs are perfectly compatible, NHEJ usually repairs the break accurately. Imprecise repair leading to loss of nucleotides can also occur but is much more common when the overhangs are not compatible. Inappropriate NHEJ can lead to translocations and telomere fusion, hallmarks of tumor cells.

TLS

Translesion synthesis (TLS) is a DNA damage tolerance process that allows the DNA replication machinery to replicate past DNA lesions such as thymine dimers or AP sites. It involves switching out regular DNA polymerases for specialized translesion polymerases (i.e. DNA polymerase IV or V, from the Y Polymerase family), often with larger active sites that can facilitate the insertion of bases opposite damaged nucleotides. Translesion synthesis polymerases often have low fidelity (high propensity to insert wrong bases) on undamaged templates relative to regular polymerases. However, many are extremely efficient at inserting

correct bases opposite specific types of damage. From a cellular perspective, risking the introduction of point mutations during translesion synthesis may be preferable to resorting to more drastic mechanisms of DNA repair, which may cause gross chromosomal aberrations or cell death. In short, the process involves specialized polymerases either bypassing or repairing lesions at locations of stalled DNA replication.

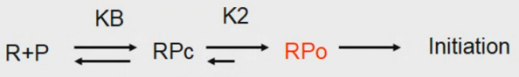
Damage Sensing Mechanisms

DNA damage needs to be sensed different sensor proteins are activated by different kind of damage, when a cell cycle checkpoint system is activated. These checkpoint systems can sense different kinds of DNA damage and employ the correct response mechanism. Sensors activate Master Kinases, in eukaryotes ATM, Chk1 and ATR, who transduce the signal to secondary kinases, CHK2, BRCA1, yH2AXm p53. These secondary kinases affect many different targets, responsible for Cell cycle arrest, DNA repair, or Apoptosis.

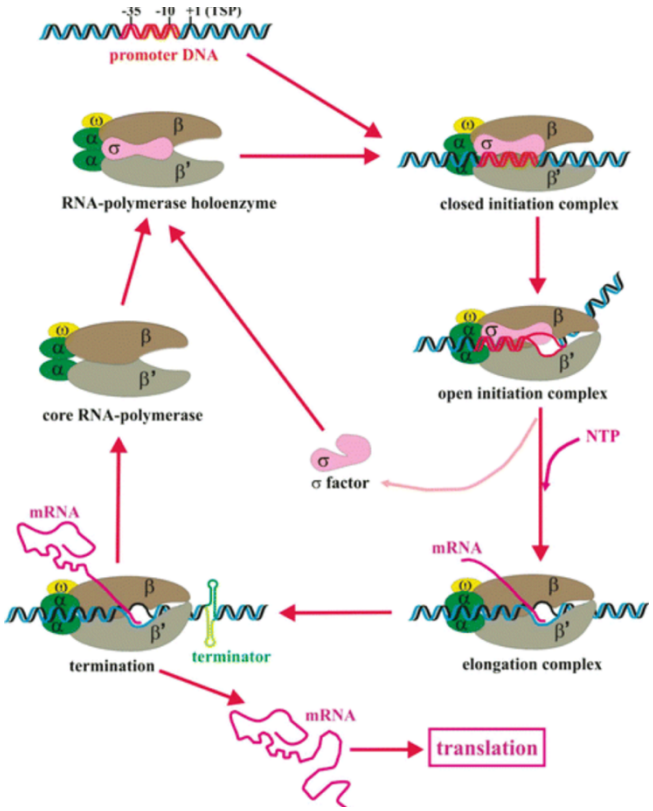
When a master kinase is recruited by a sensor, a modified single strand DNA (ssDNA) region is recognized by a specific repair pathway (NER, BER, HR) and the phosphorated master kinase initiates an activation cascade where many targets are activated.

DNA checkpoints act on Cdc25 and P53 blocking the cell cycle at the G1 or G2 checkpoints, making sure the cell does not enter cell cycle with damaged DNA. ATR will phosphorylate p53, which in turn activates many other enzymes including p21, which is a CDK inhibitor, blocking the entrance into mitosis, or the S-phase. CHK1 instead phosphorylates and inhibits Cdc25 a, b and c subunits, activators of mitosis therefor blocking mitosis in two important ways. ATM (double strand master kinase) responds in a similar way by activating p53. In cancer cells, many mutations must accumulate in order for the cell cycle control systems to be completely deactivated, showing the redundancy and fragility of the control system.

Prokaryotic Transcription:



The DNA is first denatured at the promoter, followed by the copying of one strand onto the template. Transcription and translation occur simultaneously in bacterial mRNA. This lets us define R(RNApolymerase), P(promoter) , RPc (closed complex) and RPo(open complex, promoter opened). In bacterial transcription, to calculate the speed of initiation we have two constants, KB and K2 where KB determines the speed of association of the closed complex (fast), while K2 determines the rate of transformation from the closed complex (RPc) to the open complex (RPo) (opening of the promoter, slow). This association is accelerated by the polymerases ability to jump on the DNA and scan along it until it finds a promoter, as opposed to associating and dissociating. This method allows us to understand how polymerase is able to associate with the promoter so much faster than blind-three-dimensional diffusion. KB can vary greatly, because bacteria can have good promoters, and bad promoters ranging from 10^{7}-$10^{9}/M^{-1}$. K2 instead depends greatly on KB, but varies from 10^{-3}-10^{-1}/second, meaning we form that many open promoters per second. Promoters with a higher rate constant transcribe their gene more efficiently.



Promoter sequences also have a much higher affinity for RNA Pol than other random regions of DNA, more so in prokaryotes than eukaryotes due to the large number of bases not required for transcription in eukaryotes. To

overcome this lack of increased affinity in eukaryotes, they used two major mechanisms:

1. Reduce access of competitor DNA (Nucleosomes)
2. Specificity constant increases through multiples positioning factors (GTF instead of sigma factor)

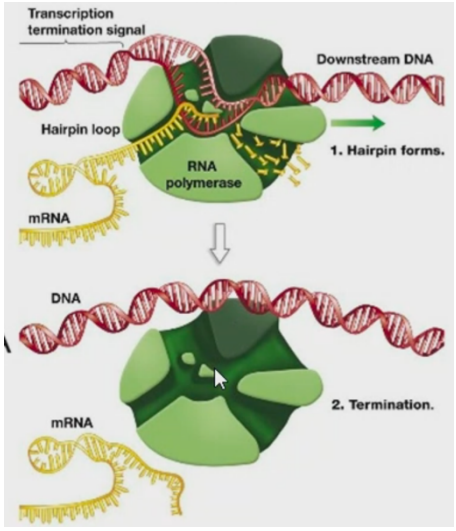
Both of which increase the specificity of promoter recognition in very large genomes --a problem that bacteria do not face due to their small genome and high concentration of regulatory DNA.

Transcription can be broken into three major phases: Initiation, elongation, and termination. Initiation can further be broken down into a series of defined steps. A promoter is a DNA sequence onto which the transcription machinery binds and initiates transcription. Although promoters vary among prokaryotic genomes, a few elements are conserved. At the -10 and -35 regions upstream of the initiation site, there are two promoter consensus sequences. The -10 consensus sequence, called the -10 region, is TATAAT. The -35 sequence, TTGACA, is recognized and bound by σ. Once this interaction is made, the subunits of the core enzyme bind to the site. The A–T-rich -10 region facilitates unwinding of the DNA template; several phosphodiester bonds are made. The transcription initiation phase ends with the production of abortive transcripts, which are polymers of approximately 10 nucleotides that are made and released, in an attempt to escape the promoter. Three models for the mechanisms of promoter escape exist, Transient passage(back and forth), Wormlike (Harmonic) , and Curling(wrapping RNA inside). We now know Curling model is correct. In all of these cases, the polymerase is likely to create abortive DNA which will not be transcribed continuously due to the anchoring, until it creates a long enough RNA to break free. The entrance into elongation phase from abortive phase is mediated by the promoter clearance, in which the RNA enters into a special channel in the polymerase, which pushes away the sigma factors, allowing the polymerase to proceeded. The general principle is that RNA elongation pushes the sigma factor out of the NT and T channels. This movement also removes binding factor, allowing growth phase.

Bacterial RNA polymerase is formed of a core made of B(chain initiation and elongation) and B'(DNA binding), plus two *a* subunits, important for assembly, chain elongation, and association with regulatory proteins. The core binds DNA but cannot recognize the promoter without the sigma subunit. This system is largely conserved over evolution, however, in eukaryotes there is absolutely no equivalent to a sigma factor. Another important factor in bacteria, called the elongation factor, cuts the RNA and allows RNA polymerase to resume transcription when it pauses in the abortive stage.

Bacterial Promoters and terminators

In the case of sigma 70, bacterial promoters have two important consensus sequences, -35, and -10, which in a good promoter are spaced by 17-19 nucleotides. When the elements are spaced out more, the promoter becomes inefficient. In some very good promoters, there is another region called the UP element, and instead binds to the alpha subunit. Another element located before -10 helps binding of the ribosome called the extended -10 element. All of these elements (except UP) are recognized by the sigma factor, different domains recognizing different sequences, while UP is recognized instead by alpha-subunit, specifically it's a-carboxy-terminal domain. UP is found in genes that are very heavily transcribed, this strong element greatly increases the transcription ability. -10, and -35 are both situated on the major groove, facing out, which is why the 17-19 is so important.



Different sigma factors are used for regulatory purposes, while sigma 70 is dubbed the general sigma factor, others are transcribed in different cellular responses. Sigma-32 for example is a cold shock factor, binding more readily to promoters necessary to respond to the cold. These alternative sigma factors hijack the polymerase to guide it to the promoters needed to respond to a specific stress. This system is largely

exploited by bacteriophages, when the cell is infected the phage transcribes its own sigma factor, substituting the bacteria general sigma factor, guiding the polymerase to the phages genes instead, also transcribing late genes that also contain sigma factors further hijacking the polymerases. This system effectively moderates the timing of the infection, the sigma factor cascade allows the cell to stay functional while the phages proteins are also being transcribed.

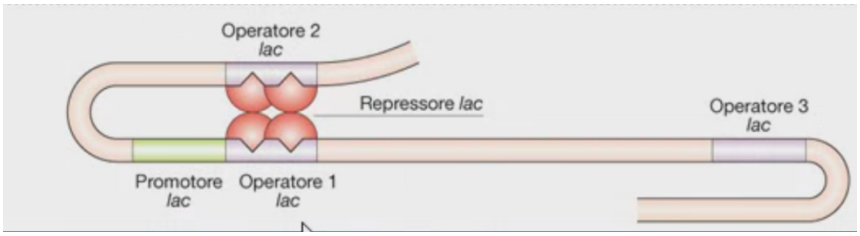
Terminator sequences, called terminators are divided into two categories:

- Rho independent terminator structures in the RNA form into a stem loop structure, followed by an Uracil rich sequence, in combination inducing termination. This happens because the stem loop structure is able to destabilize the polymerase, by provoking a steric clash in the polymerase channel, by blocking advancement. Secondly, the uracil rich sequence forms a weak DNA-RNA hybrid, forming only two hydrogen bonds with G, causing the DNA bubble to have a tendency to close.
- Rho Dependent instead not only require the hairpin structure, but also the Rho protein complex. Rho binds to the RNA upstream of the Pol., specifically to the pyrimidine rich regions, from where it slides along the RNA until it reaches Pol., when it remove Pol. Rho works by destabilizing the Pol, as well as using its helicase activity, promoting the DNA to reclose on itself, terminating transcription. Rho is formed by six subunits, which bind to RNA, and roll along until it reaches the blocked Pol.

Transcription Regulation

Although the quality of the promoter is important, promoters can also be regulated by repressors and activators, working as molecular switches to turn transcription on or off, depending on environmental conditions. Repressors work on the operator region, generally blocking the assembly of the complex, while activators work from upstream regulatory sequences. Repressors and activators can be both inductively triggered, or repressively triggered, the former being turned on by a signal, the latter being turned off by a signal.

The CAP activator protein binds in the minor groove, bending the DNA in a way the enlarges the minor groove across from the binding site, up to 120 Degrees. CAP works only when cAMP is produced, in conditions where Glu is low, signaling the need for lactase enzymes. CAP is capable of interacting with the alpha-carboxyl-terminus domain (aCTD) making the connection of the complex to the DNA more efficient, and also improving the ability for the open complex to form. The action of CAP is not only limited to anchoring the Pol., but also likely bends the DNA in a favorable way. The Lac promoter which can be activated by CAP also contains a repressor region in the promoter, called the operator.



It is now known that in addition to the operator, there are also other regulatory elements that can bind repressors. These other operator sequences can be bound by dimer repressors, binding cooperatively to each other and forming a loop which is much more efficient at repressing the transcription than only operator This looping mechanism is important in understanding how regulatory proteins can act at a distance. In bacteria, inducing curvature in certain regions of the chromosome by histone-like proteins can help facilitate the long-range regulation of promoters.

Other interesting regulation elements in bacteria involve MerR when in presence of Hg^{2+} can bind near the merT regulatory region. In normal conditions, the promoter regions -35 and -10 are too far apart to display on the major groove, but when MerR binds, a twist is induced, moving the -35 and -10 both into the major groove and yielding ideal spacing.

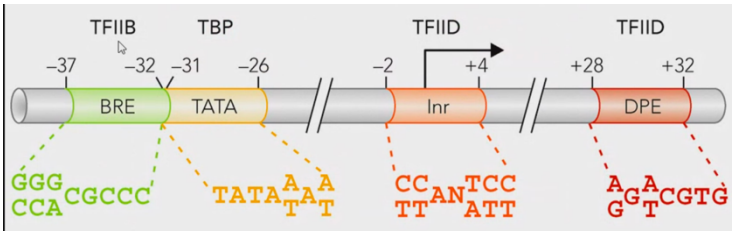
The arabinose promoter displays an example in which a regulatory protein can be an activator or a repressor. When AraC is bound to its inducer molecule, produced by arabinose being present in the body, it will act as

an activator. When instead, the AraC is not bound to its inducer, it changes confirmation to repress the promoter by forming a loop.

The trp codon is blocked when intercellular trp is high, and when low, the 2-3 hairpin blocks the operon access. Basel promoters contain some of the elements we have learned before, TATA, BRE, as well as upstream elements, and in mammals include many consensus sequences upstream called activators.

Eukaryotic Transcription

General Transcription Factors

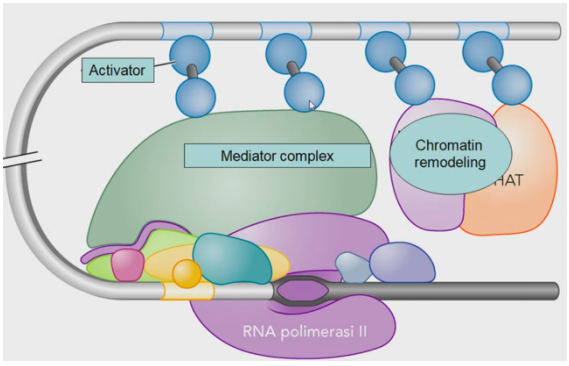


General transcription factors are the proteins that bind to the promoter sequences and establish the connection between the Pol and the promoter. These factors allow the formation of the pre-initiation complex. The basal transcription elements are:

- **TFIID,** the transcription factor devoted to recognizing the promoter sequence elements. TFIID is formed of several subunits, including the TBP, tata box binding protein. This protein binds to the tata element, when it is present (in 40% of promoters). TAF (tata associated factors) interact with the initiator, or downstream element when its present, and work similarly to the TBP.
- **TFIIA and TFIIB** stabilize the TFIID and the pre-initiation complex.
- **TFIIF** associates with RNA Pol-II and mediates the connection with D, B, and A, and the polymerase can now attach and forms the Pre-Initiation Complex
- **TFIIE and TFIIH** allow the pre-initiation complex to open the DNA on the far side of the Pol, forming an open complex, through ATP-ase activity
- **RNA Pol-II** must also be switched to the elongation mode from abortive mode, by phosphorylating the Serine 5 and Serine 2 on the tail on Pol-II, inducing a confirmation change entering elongation mode.

Eukaryotic Enhancers

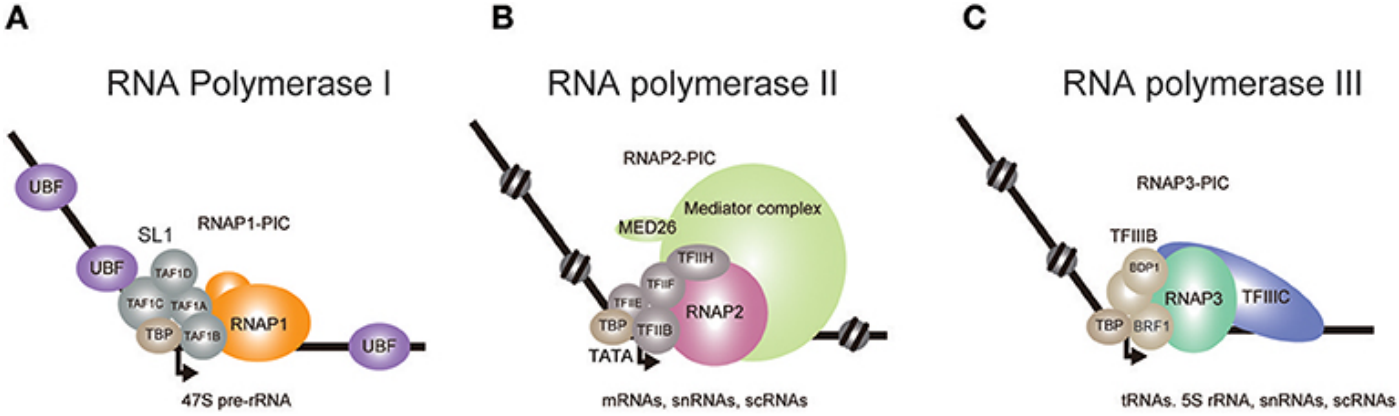
In addition to the basal promoter sequences, enhancers (UAS in bacteria) can also influence the activation of the promoter's transcription from a distance. many enhancer sequences can act on one basal promoter, and can be thousands of bases upstream, even when enhancers move around they still convey the same activation activity. This is allowed because of DNA looping, mediated by the mediator complex enzymes, when the activators are attached to the enhancer regions, which mediates the signal from the activator, as well as the remodeling of chromosome.



The first action of the activator is the binding to the enhancer, after which it can either attach to the mediator, or to the TFIID general transcription factor. When associating with the mediator, it signals either an enhanced binding stability, or a reduced ability to stabilize. Yeast mediators are made of dozens of proteins, and human mediators are even more complicated. RNA will later be capped, spliced, polyadenylated and cut, all signaled by the RNA phosphorylation signals.

RNA Pol-I is required to transcribe ribosomal RNA, which are processed to form the large and small subunits. These promoters are enhanced by UBF and SL1 binding to the upstream element, allowing for the Pol-I to associate with the basal promoter. Interestingly, the SL (selectivity factor) contains the Tata binding protein.

Pol-III transcribes ribosomal genes, and tRNA. Requires TFIIC, as well as TFIIIB, allowing for the assembly of the Pre replication complex, while the TFIIIA allows the opening of the DNA to form open complex.

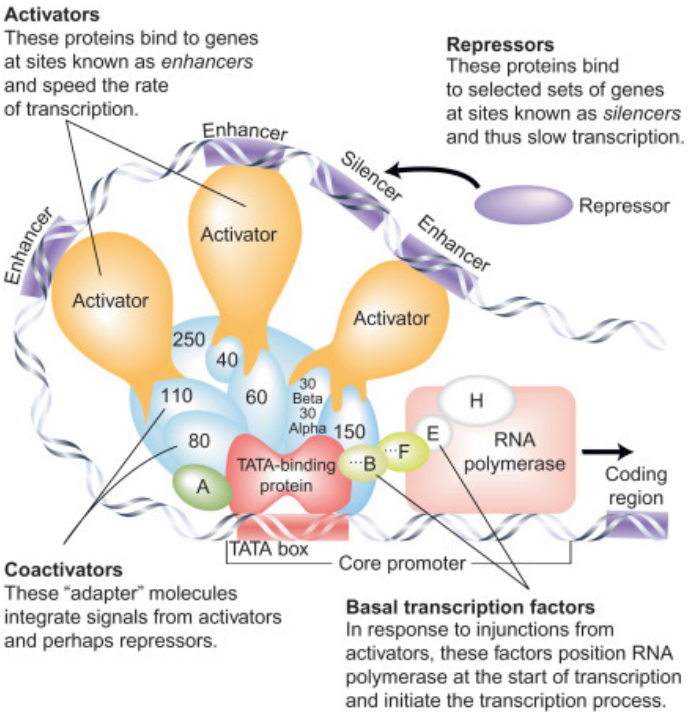


One activator protein can be switched on and off, because the binding domain and the activation domain are very distinct, the activator domain can be cut off, and the activator will not work, as it cannot contact the mediator. Likewise, the activator domain of one regulator can be added to a repressor binding domain, and the net result will be activation. Transcription factors also modulate the series of post transcriptional modifications, importantly these factors contain many different domains, each which can be moved around without affecting the others. Several domains have been diffused by the exon-shuffling theory during evolution.

Another important aspect of transcription factors is that they act in combination, some activators are stronger than others. As mentioned above, some activators can compete with nucleosomes for positioning at promoters, opening the space for transcription. They can also recruit histone acetyl transferases, which can acetylate histones, relaxing chromatin and increasing transcription. They can also recruit Chromatin remodeling complexes, which can move or disassociate nucleosomes, allowing for formation of transcriptional complexes.

Insulators and Repressors

Insulators are another DNA binding element that blocks the activating action of the enhancer, by somehow blocking the signals. Insulators can work for upstream elements, downstream, or both when they are found on both sides of the promoter. The mechanism is still not fully known how the insulator works, but likely it blocks the enhancer bound proteins from contacting the promoter region.



In eukaryotes, there are many mechanisms of repressions:

1. Competitive inhibition: the inhibitor blocks the activator from binding
2. Direct inhibition: The inhibitor blocks the activator by binding to it such that it can't bind the mediator/promoter proteins
3. Direct repression: The repressor destabilizes the mediator directly, making it harder for transcription factors to stabilize and fire
4. Indirect repression: The deacetylation of the histones by repressors that bind the histones, tightening the chromatin and making it harder to transcribe

Pervasive Transcription

Cells receive signals from the environment, which are transduced in various pathways, one of which activates transcription factors. In the Ras-MAP Kinase pathway the transcriptional activator is phosphorylated by MAPK and can then enter the nucleus.

In pervasive transcription, it seems the whole genome including introns, snRNAs, miRNA, and snaRNA is all transcribed. Most of these unwanted RNAs are degraded by the exosome enzyme. Pervasive transcription is now accepted to be a general feature of eukaryotic genomes, generating short and long non-coding RNAs (ncRNAs). Growing number of examples have shown that regulatory ncRNAs can control gene expression and chromatin domain formation. Other RNA such as CUTs (cryptic unstable transcripts) and Prompts, increase in the absence of nucleases. In summary, transcription is not exclusive to the mRNA, but short strands are prematurely aborted, as well as being degraded by exonucleases.

Other RNAs are stable and serve as functional units, such as Scaffold functions, Guiding RNA-Pol tethering, mediating long range interactions by forming loop in DNA, transcriptional interference in the promoter, or antisense destabilization, or interfering with nucleosomes. These are only some of the way's RNA can affect the cell in a catalytic fashion.

Lambda Phage growth

Lambdas inoculate their genome inside a bacteria, where is proceeds either to integration with the bacterial chromosome, reproducing when the cell reproduces(lysogenic route) or the genome replicates in many copies as fast as possible promoting ligase of the cell and distribution of the phages (induction route). Bacteria exposed to UV light in small amounts, the phage will favor the induction route, the lambda knows the cell will die so it wants to leave the genome before they both die.

This mechanism of regulated by Rec A protein, a proteolytic enzyme that degrades the lambda repressor in the presence of UV, and the cell enters the induction route. Another scenario which lytic growth is favored over lysogenic is when the cell is multiplying very fast, therefor there are many bacteria the phage can infect. In this case, the CII protein promotes the transcription of the lambda repressor, is degraded by the lambda phages and the lytic growth is preferred.

The mechanisms of Lambda repressor are well studied. Lambda repressor blocks the transcription of Cro gene and later genes, while also promoting the transcription of itself. In antitermination the terminator sequence promotes premature termination of bacterial genes unless Lambda Q protein is bound.

Techniques of Molecular Biology:

FGS (First Generation Sequencing)

DNA sequences can be read using special gel electrophoreses, but the DNA must be cut, labeled and separated, as well as having to read the output manually, making it an extremely time-consuming process. Another method was later proposed in which a DNA Pol starting from a dimer would copy a length of DNA using labeled nucleotides molecularly modified called Chain Terminators to block addition of another base after attached. These chain terminators need to be mixed in correct proportions in order to get enough reading while still allowing the chain to elongate. One of the four nucleotides must be labeled so it will appear in the gel. In modern labs, four different fluorochromes are used to distinguish the bases.

NGS

New generation techniques instead fragment and sequence without closing, increasing the efficiency from 700 to billions sequences read per experiment. Present sequencing called third gene sequencing allows us to work

with single DNA molecules, without the need to fragment. This has drastically decreased the cost of DNA sequencing, allowing much more data to be available.

DNA sequencing with commercially available NGS platforms is generally conducted with the following steps. First, DNA sequencing libraries are generated by clonal amplification by PCR in vitro. Second, the DNA is sequenced by synthesis, such that the DNA sequence is determined by the addition of nucleotides to the complementary strand rather than through chain-termination chemistry. Third, the spatially segregated, amplified DNA templates are sequenced simultaneously in a massively parallel fashion without the requirement for a physical separation step.

PCR

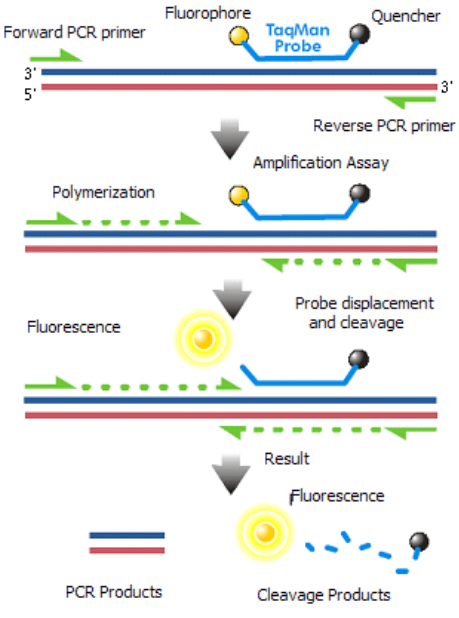
Polymerase chain reaction (PCR) is a method widely used to rapidly make millions to billions of copies of a specific DNA sample, allowing scientists to take a very small sample of DNA and amplify it to a large enough amount to study in detail. PCR generally requires two reagents; primers (Used as markers for the DNA) and polymerases (Taq polymerase from thermophilic bacteria, which is heat stable). Primer design is an important area of research. Primers must be complementary to a region on the target strand, but they do not necessarily have to be a 100% complement, and therefor can be used as biomarkers. Primer length is also important -short primers are used mainly for amplifying small fragments of DNA, while long primers are used to amplify eukaryotic genomic samples. The heat resistance is necessary as a process called heat cycling is used, the fluctuation from low to high heat allows all the reaction pathways to proceed. The basic steps are:

1) Denaturing(96°): Heat the reactant to separate the two DNA strands, providing ssDNA templates
2) Annealing(55-65°): Cool reaction, so primers can bind to their complementary sequences of ssDNA
3) Extension(72°): Increase temperature, so Taq polymerase extends the primers, synthesizing target DNA

The cycle repeats 20-30 times in typical PCR, taking around 2-4 hours depending on the reaction efficiency. In this time, billions of copies of the region can be made, because Taq polymerase also replicates each replicated strand (and its primers) as well as the originals. It is routinely used in DNA cloning, medical diagnostics, and forensic analysis of DNA.

RT-PCR

Reverse transcription polymerase chain reaction (RT-PCR) is a laboratory technique combining reverse transcription of RNA into DNA and amplification of specific DNA targets using polymerase chain reaction (PCR). It is primarily used to measure the amount of a specific RNA. This is achieved by monitoring the amplification reaction using fluorescence, a technique called real-time PCR or quantitative PCR (qPCR). Combined RT-PCR and qPCR are routinely used for analysis of gene expression and quantification of viral RNA in research and clinical settings.



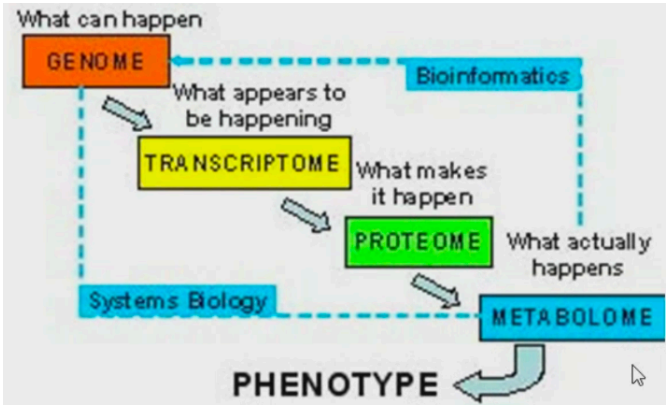
Real-time RT-PCR relies on novel florescent DNA labeling techniques, allowing analysis and detection of PCR products in real time. Not only is real-time RT-PCR now the method of choice for quantification of gene expression, but it is also the preferred method of obtaining results from array analyses and gene expressions on a global scale. Different probes are used, such as SYBR green, or the one we will discuss TaqMan probe, but their job it always to serve as a florescent beacon.

TaqMan probes are oligonucleotides that have a fluorescent probe attached to the 5' end and a quencher to the 3' end. During PCR amplification, these probes will hybridize to the target sequences located in the amplicon (The region of DNA we want to replicate) and as

polymerase replicates the template with TaqMan bound, it also cleaves the fluorescent probe due to polymerase 5'- nuclease activity. Because the close proximity between the quench molecule and the fluorescent probe normally prevents fluorescence from being detected through FRET (Resonance energy transfer), the decoupling results in the increase of intensity of fluorescence proportional to the number of the probe cleavage cycles.

Gene regulation and proteomics

The least wasteful choice for regulation is transcriptional regulation, where no cell materials are 'wasted'. There is however other important mechanism of regulation, deemed post-transcriptional, and post translational regulations which generally act faster than transcriptional regulation. Post transcriptional regulation specifically focuses on the regulation of stability, transport, maturation, and translation of mRNA, while post translational regulations focus on turning on or turning off the activity of the protein, even permanently through proteolysis.



The most powerful way of studying all of these influences is still through the study of RNA. For bioinformatics, this means studying metabolome, the fluctuation in phenotype. While the metabolome (Metabolic molecules in use) is often hard to study in a lab, the proteome (all possible proteins in the genome) can be easier to study but is still less global than transcriptome (all mRNA and tRNA in the cell). For this reason, we often choose to study transcriptomics for predicting phenotypes due to its global nature. Because of the possible regulation steps such as post-transcriptions and post-translational, the results from the transcriptome may not necessarily correlate to what we would expect, that is the mRNA is not guaranteed to translate to a protein, nor can we assume a protein that is created will be functional causing a change in metabolome.

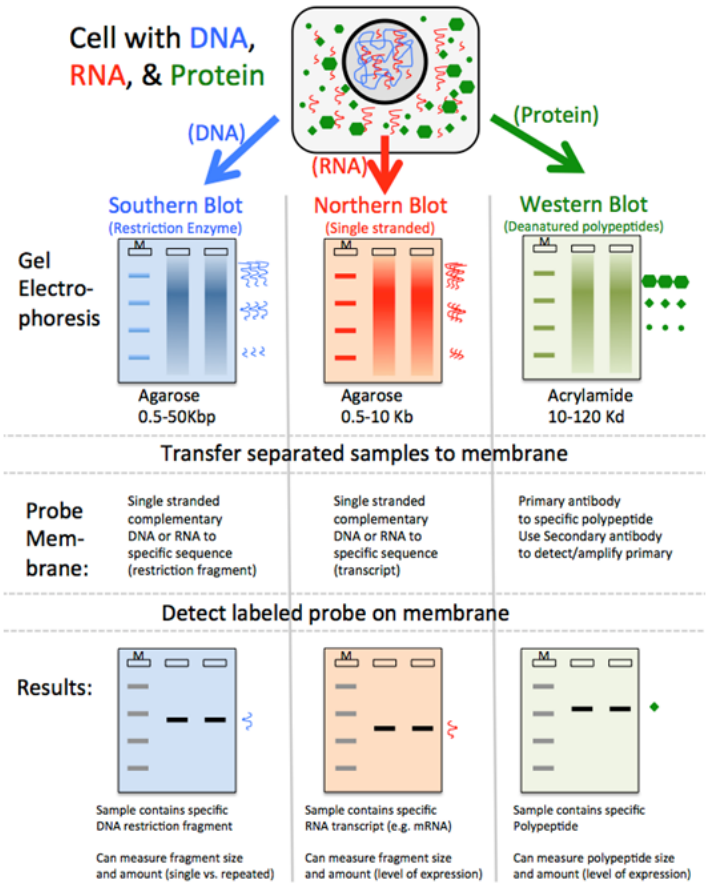
Proteomics has evolved from gel electrophoresis, to so called high-throughput proteomics. Proteins HTP are digested into a peptide mixture, which is introduced into a mass spectrometry/gas analysis, from which the peptides can be identified as peaks in the output and match them to elements in a database. In reality, because of the massive number of proteins present in higher organisms, the peaks are too small to properly identify a protein.

For this reason, transcriptomics is more often studied, in which DNA microarrays are synthesized in a way that we can visualize each RNA that will be synthesized in each pane of the microarray. These microarrays let us study gene regulation, gene expression changes, or even diagnostic microarrays used as biomarkers for sick vs healthy tissue. They can also be used to visualize genome organization, chromatin organization and mutations. The best sequences to assemble microarrays is by using very short gene sequences, not always possible.

Reverse transcription can be used, along with DNA denaturation in such a way that the singe stranded DNA binds to the reversely transcribed elements. These double stranded regions are much more stable than the single stranded regions, and have a much better resistance to denaturation, allowing the target regions of DNA to be separated.

Blot tests

Southern blotting is one of the oldest techniques used in which DNA is fragmented and separated using gel-electrophoresis. Once all segments of substrate are displayed on the gel, probes are implemented which chemically mark the substrate in a way that allows the targeted substrate to be detected. Blot tests but be taken with special consideration due to the post transcriptional modifications, the quantification of the generic material cannot guarantee a result. To overcome this, we can measure a known gene that does not change in the conditions and use it to calibrate the change in substrate. This calibration can be negative or positive and is an important step in interpreting results. Blot tests look closely at a specific gene, while transcriptomics studies all the genes.



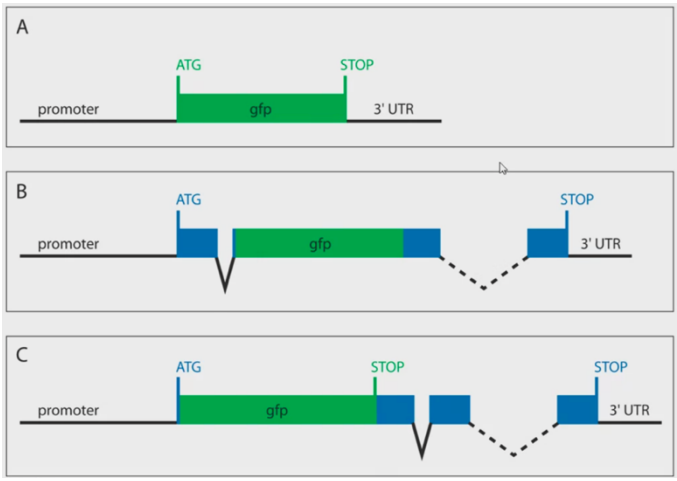
Nuclear Run-on is another technique used to mark RNA with radioactive nucleotides. We can then study the increase in transcription of a specific RNA, when increasing we can say the 'on rate' is increasing as opposed to the stability is increasing.

To study stability of RNA, we simply block transcription by deactivating RNA Pol. Stable RNA will show a slow decay curve, while hugely unstable RNA is decayed rapidly. These processes are called RNA decay curves.

Western Blot measures protein concentration based on a specific antibody binding to a protein on a filter, in this way the blots appear on the lane of the conjugate antibody. In this case we need to use a calibration antibody for each lane, that selects for a protein other than the one studying, and the blot for this calibrator antibody must be the same in every lane, or there was selection bias. The calibrator can also be used to interpret results -always assuming the calibrator protein density is constant.

Reporter Genes

Reporter genes are able to easily quantify the gene expression of an area of DNA. In this methodology, we have to find a reporter gene that is activated by the same promoter as the gene we are studying. They give us interesting information about how the promoter responds to changing environments in vivo. In the lab, the coding region for the reporter gene must be inserted into the coding region after the promoter in question(A) to study the promoter. To study the stability of the mRNA, the reporter gene must be inserted into the gene coding region without disturbing the mRNA(C), and lastly we can study the localization of the protein(B) by inserting the reporter gene after the localization signal.



EMSA

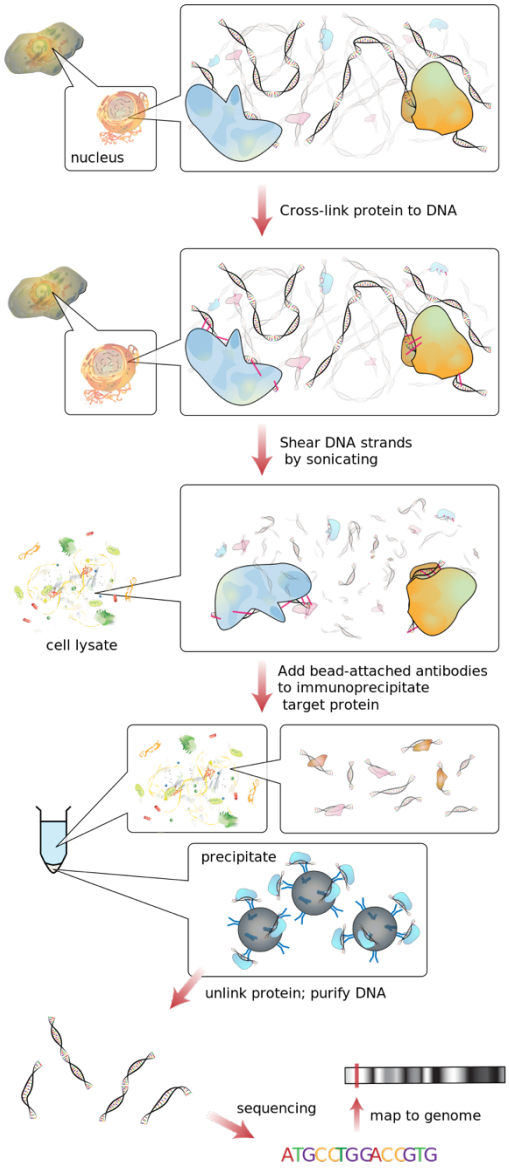
EMSA (Electrophoretic mobility shift assay) is the simplest methodology for studying proteins that bind to DNA. In this approach, a median with the protein is inserted with possible DNA segments, the ones that bind to the protein in the gel will move more slowly through the gel. DNA segments are run through different lanes; therefore, we can also estimate the approximant binding rate of the protein to the DNA. This interaction could be non-specific, in which case all lanes will show the decreased movement speed through the gel, while if its specific, some DNA segments will not slow down, and others will.

Footprint

If we are unsure of which region of DNA is binding to the specific protein, we have to run another test called a footprint test. In foot printing, small segments of DNA are cut into short segments, marked, and tracked in lanes. In -OH (hydroxy radical) foot printing the nucleophile attacks all the bases, cutting them into small, even pieces while in DNase 1 foot printing the enzyme DNase cuts the DNA at various locations, making -OH much more accurate and specific for indicating a region which the protein binds.

Chromatin Immunoprecipitation (ChIP)

ChIP is the protocol that allows us to study the localization on the genome of gene regulatory proteins in vivo. The protocol is: Crosslinking → Sonication → Immunoprecipitation → Reverse crosslinking → DNA extraction and analysis. We can map either 1) the protein that binds to a specific site on the genome, or 2) map all the possible binding locations of a specific protein on the genome. The second is especially important for bioinformatics.



Crosslinking involves the locating of all the proteins attached to the genome, and fixing their location, so that when the structure is broken to study, the proteins remain attached to the DNA. This involves specifically the conversion of the weak-noncovalent bonds between protein-DNA to covalent bonds. The most frequent way to do this is by adding formaldehyde to the cell, a crosslinking agent that forms covalent bonds between the protein-DNA, but also between protein-protein subunits.

Secondly, we break the nucleus of the cell, and extract the chromatin. The breaking of the nuclei varies depending on the cell type. We must then however break the chromatin into fragments, which id don't at random by sonication. In sonication, a probe emitting ultrasound waves is inserted into the immediate area of the chromatin, breaking the chromatin into 300-1000 base pair pieces. We aim for 500 bp subunits through trial and error, breaking and then doing an electrophoresis test with the fragments to find the average size.

Finally, in immunoprecipitation we take an antibody that recognizes a certain element on the protein we want to study, mark them and insert them into the mixture, and they will bind to the DNA-protein complex. Lastly, a bead with an antibody binding property is inserted, marking all antibodies that are bound. We cannot separate the protein-DNA complexes we wish to study from the other and heat the final solutions to remove the crosslinks. Proteases are then inserted to digest the DNA binding proteins, and just the DNA regions with the protein binding sequence remain. These DNA sequences must then be analyzed in order to find the exact

sequence related to protein binding. ChiP-Seq is a powerful tool to study proteins with known domains, as well as important protein-DNA interactions.

CRISPR

CRISPR (clustered regularly interspaced short palindromic repeats) is a family of DNA sequences found in the genomes of prokaryotic organisms such as bacteria and archaea. These sequences are derived from DNA fragments of bacteriophages that had previously infected the prokaryote. They are used to detect and destroy DNA from similar bacteriophages during subsequent infections. Cas9 (or "CRISPR-associated protein 9") is an enzyme that uses CRISPR sequences as a guide to recognize and cleave specific strands of DNA that are complementary to the CRISPR sequence. Cas9 enzymes together with CRISPR sequences form the basis of a technology known as CRISPR-Cas9 that can be used to edit genes within organisms. CRISPR technology has been applied in the food and farming industries to engineer probiotic cultures and to immunize industrial cultures (for yogurt, for instance) versus infections. It is also being used in crops to enhance yield, drought tolerance and nutritional homes

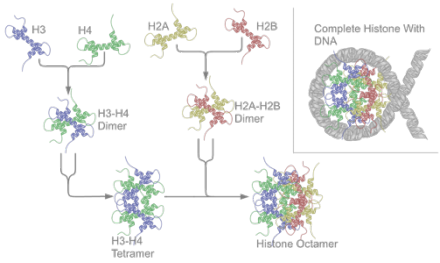
Chromatin

Chromatin was originally thought to be the means to compact and order the genetic material. More recently, we have discovered the important role it plays in epigenetic regulation, as well as gene expression. Chromatin has many levels of compaction, the first is represented by the beads on a string model, with a packing ration of 6-7 folds, compared to free DNA with packing ratio of 1. Secondly, the solenoid structure which reaches 40-fold packing. The major compacting, however, comes from the formation of loops, typical of interphase and a further looped structure in mitotic chromosomes. These loops are anchored to the matrix, with up to 680 packing ratio.

Nucleosomes inhibit accessibility to promoter and binding of transcription factors to regulate sequences. This implies active promoters are generally nucleosome dependent, as well as gene transcription activation is coupled to nucleosome eviction/displacement, and nucleosome occupancy in promoters is anti-correlated with transcription activity level. In summary, nucleosomes obstruct transcription machinery, and chromosome remodeling, histone modification, and substitution of canonical histones with variant isoforms are strategies used to remove these obstacles.

Certain regions of the chromatin are hyperactively transcribed compared to the rest of the genome; these active nuclear compartments are decondensed chromatin loops near the nuclear pours which can be easily transported out of the nucleus after maturation.

Histones



Histones are characterized by a global portion, and some linear tails. Core histones are H2A, H2B, H3 and H4, and are highly conserved small basic proteins formed by 3 alpha helixes creating the 'histone fold' region of the protein. The globular and helix regions are very well conserved, while the tails are less conserved. The tales are also positively charged due to their highly basic amino acid content, especially lysine and arginine. In addition to the four core histones mentioned above, there are also linker histones such as H1 which play an important role in connecting histones, particularly because their tails contain amino acids which can be easily phosphorylated.

Histones form an octamer by first forming a H3-H4 tetramer, as well as a H2A-H2B dimer and covalently linking them together. The octamer is able to form the beads on the string structure, but the next compaction is the 30-nm fiber, formed by six nucleosomes per super helical turn. Linker proteins play an important role in forming the solenoid structure from the beads-on-a-string nucleosome array. These next level packing methods reduce the ability of the associated DNA regions to be transcribed, while the less condensed regions are more likely to be transcribed.

The structure of the nucleosome is very well studied, the difficulty was determining the location of the tales. Nucleosome DNA is 147 bp long, with 14 strong point of interaction between the basic amino acids of the histones H3 and H4 with DNA, meaning there is about an interaction every 10 base pairs. This interaction is on the minor groove, facing inward in a way that compresses the minor groove, increasing the size of the major groove on the other side of the DNA molecule. These interactions are mostly between the positively charged histidine amino acids, with the positively charged phosphate groups on the DNA. This electrostatic interaction implies that the sequence of DNA is irrelevant, as all base pairs contain a phosphate group. This is obviously not the case, but the reason is due to flexibility of the DNA regions, the DNA in the region interacting with the nucleosome must be adequately flexible to be wrapped, too rigid and its roll angel will not be sufficient to wrap around the histone. AA/TT regions do not allow bending and are found every 10 base pairs away from the histone-DNA interactions, while GC/CG sequences are found near the histone-DNA interaction because they are bendable, therefor are found in the regions of the major groove that need to be opened.

Flexibility however does not tell the whole story, for example protein complexes, or unusual DNA patters such as crucifixes can block chromatin formation, thus not allowing the nucleosome structure to assemble there. In this case, the nucleosome will move on until it finds a desirable/available sequence. This mechanism is called the boundary mechanism and can be summarized as: if the nucleosome cannot access the normal binding region due to it being occupied, or otherwise unusual, it will instead bind to the boundary of this blockage, at the first available position. Flexibility can be used as the primary predictor for nucleosome occupancy, but when the vivo binding pattern varies from the expected binding patters, we can infer that a protein complex can compete for those sequences and prevent nucleosomes from binding.

In all, histones make five types of interactions with DNA:

- Helix-dipoles form alpha-helixes in H2B, H3, and H4 cause a net positive charge to accumulate at the point of interaction with negatively charged phosphate groups on DNA
- Hydrogen bonds between the DNA backbone and the amide group on the main chain of histone proteins
- Nonpolar interactions between the histone and deoxyribose sugars on DNA
- Salt bridges and hydrogen bonds between side chains of basic amino acids (especially lysine and arginine) and phosphate oxygens on DNA
- Non-specific minor groove insertions of the H3 and H2B N-terminal tails into two minor grooves each on the DNA molecule

From bioinformatical analysis we can find two situations for promoters. In type-II architecture, we find a nucleosome free region on the promoter through competitive binding of another protein, this is the case of housekeeper proteins which should almost never be inhibited by nucleosome formation. In type-I architecture, the proteins which compete with the nucleosome can either be bound or unbound, typically for promoters of genes that are highly regulated. Type-II regions containing large sections (500bp) of nucleosome depleted DNA especially at the origin of replication. Type-II promoters make up about 630 genes that are depleted of nucleosomes, while Type-I makes up 390 genes, and its nucleosome concentration varies greatly. These occupancies are studied particularly on the regions directly upstream of the region start, the so called -1 nucleosome. This -1 region is mostly NSR (nucleosome free region) in housekeeping genes and is indicative of the transcribability of the entire coding region. High nucleosome occupancy is therefore associated with low mRNA transcription, but less obviously leads to a higher level of transcriptional plasticity, the ability to

regulate the amount of transcription as well as higher sensitivity to histone regulation. Nucleosome organization around promoters is evolutionarily conserved, indicating the important role nucleosomes play in gene regulation, and epigenetics.

Histones can be remodeled in three ways: sliding, transferring, or remodeling (opening without moving the disc). These movements can allow or block transcription. Histones dimers can also be substituted with histone variants, which change their activity (discussed below). These changes, as well as the inheritance of histones is modulated by special histone chaperone proteins, one of which (CAF-1) is able to interact with the sliding clamp protein at the replication fork.

Some histone variants are used for special processes in the body, others are caused by mutation. H2A.X is present in 10-15% of nucleosomes, and plays a crucial role in DNA damage response concerning double strand breaks. CENP is important which occurs in centromeres in order to promote kinetochore assembly. H3.3 and CID replace H3 in embryotic development can lead to transcriptional reprogramming, and methylation properties. H2A.Z replaces H2A, changing the confirmation of the chromatin into a conical structure. H2AZ is found in promoters which need to be very rapidly switched on, a process which is mediated by protein complex that modifies the histone immediately before the promoter.

Histone Modification

Histone tails can be modified in order to close or open chromatin. These tails are the elements that are accessible by histone modifying protein. It has been shown that experimentally, when nucleosome tails are cleaved artificially the nucleosome becomes more mobile on the genome.

Lysine methylation: The addition of one, two, or many methyl groups to lysine has little effect on the chemistry of the histone; methylation leaves the charge of the lysine intact and adds a minimal number of atoms so steric interactions are mostly unaffected. However, proteins containing Tudor, chromo or PHD domains, amongst others, can recognize lysine methylation with exquisite sensitivity and differentiate mono, di and tri-methyl lysine, to the extent that, for some lysine's (e.g.: H4K20) mono, di and tri-methylation appear to have different meanings. Because of this, lysine methylation tends to be a very informative mark and dominates the known histone modification functions.

Arginine methylation: What was said above of the chemistry of lysine methylation also applies to arginine methylation, and some protein domains—e.g., Tudor domains—can be specific for methyl arginine instead of methyl lysine. Arginine is known to be mono- or di-methylated, and methylation can be symmetric or asymmetric, potentially with different meanings.

Lysine acetylation: Addition of an acetyl group has a major chemical effect on lysine as it neutralizes the positive charge. This reduces electrostatic attraction between the histone and the negatively charged DNA backbone, loosening the chromatin structure; highly acetylated histones form more accessible chromatin and tend to be associated with active transcription. Lysine acetylation appears to be less precise in meaning than methylation, in that histone acetyltransferases tend to act on more than one lysine; presumably this reflects the need to alter multiple lysine's to have a significant effect on chromatin structure.

Serine/threonine/tyrosine phosphorylation: Addition of a negatively charged phosphate group can lead to major changes in protein structure, leading to the well-characterized role of phosphorylation in controlling protein function. It is not clear what structural implications histone phosphorylation has, but histone phosphorylation has clear functions as a post-translational modification, and binding domains.

Many histone methylation enzymes exist, mostly regulation transcription. Some regions of the genome, called heterochromatin, are tightly, permanently packed regions which cannot be methylated or acetylated. Histone phosphorylation instead is mainly involved in DNA repair, DNA damage sensing and apoptosis. Sumoylation and ubiquitination are other covalent modification on histones, mostly involved in gene expression and DNA

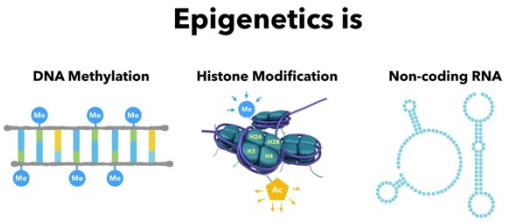
repair, respectively. Ubiquitination is not marking the histone for degradation in this case, instead it's another way of marking the histone for potential restructuring.

The so called 'histone code' is composed of many messages including the five mentioned above, together signaling if the histone should be relaxed, or increasingly tightened. This code is read by special enzymes called 'writers' which modify the histone tails, such as acetylases, or methylases. Other enzymes, called 'erasers' can undo these modifications, such as deacetylases, and demethylases. These modified histone tails are recognized by writer proteins containing a bromo domain, which recognizes the modified tails and recruits other writers with its HAT domain. One example of the HAT proteins is TAFII-250, which has two bromo domains which is specialized for acetylation. Cromo domains instead recognize and modify methylation elements in the same way as bromo recognizes and modifies acetylation elements.

These writers are aided by proteins called nucleosome remodeling complexes, which are the ATP driven proteins which actually do the opening or tightening of the chromatin. Nucleosome remodeling complexes achieve this through sliding, transferring, or remodeling (by bending or wrapping alternatively) in such a way that the associated DNA is either more or less accessible. Most of the time the remodeling favors transcription. Certain polycombs complexes prepare very repressed chromatin characterized by tri methylation of chromatin lysine 3, which recruits factors to extensively prevent transcription, as well as blocking acetylation.

Many histone modification are inheritable by offspring, meaning regions that are being tightly compressed will continue to be repressed. This is epigenetic because the traits are not inherited because of a difference in the DNA, but the inheritance patters are still conserved. The replication of histones is semi conservative -like DNA. This means histones are incorporated where they are missing on the daughter DNA strand. These histone proteins are added by special proteins but later the histone tails are modified covalently, using the old DNA strand as a template.

Epigenetics



Epigenetics is the study of heritable phenotype changes that do not involve alterations in the DNA sequence. DNA methylation is an epigenetic mechanism that occurs by the addition of a methyl (CH_3) group to DNA, thereby often modifying the function of the genes and affecting gene expression. The most widely characterized DNA methylation process is the covalent addition of the methyl group at the 5-carbon of the cytosine ring resulting in 5-methylcytosine (5-mC), also informally known as the "fifth base" of DNA. These methyl groups project into the major groove of DNA and inhibit transcription. Methylation can change the activity of a DNA segment without changing the sequence. When located in a gene promoter, DNA methylation typically acts to repress gene transcription. In mammals, DNA methylation is essential for normal development and is associated with a number of key processes including genomic imprinting, X-chromosome inactivation, repression of transposable elements, aging, and carcinogenesis. This mechanism enables differentiated cells in a multicellular organism to express only the genes that are necessary for their own activity. Epigenetic changes are preserved when cells divide. Most epigenetic changes only occur within the course of one individual organism's lifetime; however, these epigenetic changes can be transmitted to the organism's offspring through a process called transgenerational epigenetic inheritance.

The second covalent modification effecting epigenetics is the post translational modification of the amino acids that make up histone proteins. Histone proteins are made up of long chains of amino acids. If the amino acids that are in the chain are changed, the shape of the histone might be modified. DNA is not completely unwound during replication. It is possible, then, that the modified histones may be carried into each new copy of the DNA. Not all histones modifications are epigenetically, only 10-20% are retained.

Small RNAs (sRNA) is also important in epigenetic modification. In the sperm cells, circular RNA and sRNA are relatively stable, acting as translation regulators that are conserved. Other epigenetically regulations come from diet, drug use, and long-term mental health issues.

Epigenetic research uses a wide range of molecular biological techniques to further understand of epigenetic phenomena, including chromatin immunoprecipitation (together with its large-scale variants ChIP-on-chip and ChIP-Seq), fluorescent in situ hybridization, methylation-sensitive restriction enzymes, DNA adenine methyltransferase identification (DamID) and bisulfite sequencing. Furthermore, the use of bioinformatics methods has a role in computational epigenetics.