Biology 367 Fall 2016

Selected Concepts and Facts From Biology 367

Background Information
These notes present some of the facts and concepts covered in Biol 367. **They are meant to be a reviewing aid and are not meant to cover all the materials presented in Biology 367**.

The molecular nature of genes (Chapter 2)

The genetic material DNA is the nearly universal genetic material The first experimental evidence for this was obtained by showing that DNA caused bacterial transformation (Griffiths experiments which were extended by MacLeod, McCarty and Avery). These results were verified by experiments of Hershey and Martha Chase showing that DNA not protein was the macromolecule responsible for virus production in phage-infected bacteria.

DNA is a double helix with two antiparallel strands According to the Watson and Crick model proposed in 1953 the DNA molecule is a double helix composed of two antiparallel strands of nucleotides, each nucleotide consists of one of four nitrogenous bases (A, T, G or C), a deoxyribose sugar and a phosphate. An A on one strand pairs with a T on the other, and a G pairs with a C. The sequence of bases in the DNA of an organism codes for the information required for an organism's replication. The information that DNA codes for is specified by the sequence of its bases.

DNA Replication is semiconservative The DNA molecule reproduces by semiconservative replication. In this type of replication the two DNA strands separate and the cellular machinery then synthesizes a complementary strand for each. By producing nearly exact copies (the replication machinery of most organisms make less that one mistake in every billion base pairs synthesized) the base sequence information in DNA is accurately replicated thereby allowing life to replicate itself.

Melting temperature of DNA DNA is a double helical molecule with each single strand held to its compliment by weak hydrogen bonds. Because of the weak nature of these hydrogen bonds, double stranded DNA molecules can be easily converted into single stranded molecules by heating the DNA above its melting temperature. The Tm (melting temperature) for a solution of double stranded DNA is the temperature at which half of the DNA is converted to its single stranded form. The Tm for DNA is dependent upon its base composition. For example, the Tm for DNA with only A-T base pairs is about 48C while the Tm for DNA with only C-G base pairs is about 83C (this is in a weak salt (0.01M NaCl) solution. Of course the AT content of naturally occurring DNA is never 100% or 0% but some intermediate value.

Note: The Tm is dependent upon the AT content. This is the same as saying that the Tm is dependent upon the CG content.

Base pair specific hydrogen bonding holds the complementary strands together The two strands of a DNA molecule are held together by complementary base pairing (remember the two strands of a DNA molecule are held together by base specific hydrogen bonds that form G-C and

A-T base pairs). When in its double stranded form the two strands of a DNA molecule align in an anti-parallel fashion.

Denatured DNA can re-associate but reassociation requires complementary strands If you denature (convert to its single stranded form) the DNA in a solution by heating it above its Tm you can get it to reassociate or renature (return to its double stranded state) by cooling the solution of single stranded molecules to below the Tm and waiting for the complementary strands to reassociate. Note: The single stranded DNA molecules can only reassociate with their complementary strands. This is because of the very specific nature of the hydrogen bonding in DNA. For example, the sequence 5'-GGATTTCT-3' only reassociates (form a double stranded molecule) by hydrogen bonding with the sequence 5'-AGAAATCC-3'.

How quickly denatured DNA renatures depends upon the concentration of complementary strands The renaturation rate of DNA depends on the concentration of each type of complementary DNA fragment. That is, the collisions necessary for complementary single stranded molecules to interact so that they can renature depends on the concentration of both complementary strands. If the strands are present in high concentrations, then successful collisions are more likely than if the strands are present in low concentrations.

The renaturation rate of two complementary DNA molecules, S and its complement S', is proportional to the concentration of S and S' (or the rate of reassociation is proportional to [S][S']). If you know the rate at which a DNA solution reassociates you can estimate the concentration of the single strands that are reassociating.

Renaturation and DNA fragment size Renaturation rate is affected by the size of the single-stranded DNA fragments. If you want to compare the renaturation rates of two different DNA samples it is therefore necessary to shear (break into small fragments) the DNA samples so that the fragment sizes are the same.

Gel electrophoresis of nucleic acids The size of DNA and RNA fragments can be determined by gel electrophoresis. DNA and RNA fragments come in sizes ranging from a few bases to kilo bases to hundreds of mega bases (DNA only). In lectures various types of gel electrophoresis were described (denaturing gels vs nondenaturing gels, agarose vs polyacrylamide, and standard electrophoresis and pulsed-field gel electrophoresis). These various electrophoretic methods enable researchers to resolve DNA molecules varying in size from a few bases or base pairs to mega bases in length.

An introduction to gene function (Chapter 3)

Most genes code for proteins A small fraction code for RNA species like tRNAs and rRNAs. Proteins are polymers of amino acids linked through peptide bonds. The sequence of amino acids in a polypeptide (primary structure) gives rise to that molecule's local shape (secondary structure for example regions of alpha-helix, beta-pleated sheet and turns) and overall shape (tertiary structure). The primary sequence ultimately also determines whether or not a peptide will interact with other polypeptides. The highest level of structure results from the interaction of more than one peptide to generate multisubunit proteins. This level of organization is referred to as the quaternary structure.

Gene expression Gene expression is the process by which cells convert the DNA sequence of a

gene to the RNA sequence of a transcript, and then decode the RNA sequence to the amino acid sequence of a polypeptide. Ribosomes with the aid of tRNAs synthesize peptides using mRNA as the template.

How DNA codes for protein The nearly universal genetic code consists of 64 codons composed of three nucleotides each. Of these codons, 61 specify amino acids, while 3 (UAA, UAG and UGA) are nonsense or stop codons that do not specify an amino acid. The code is degenerate: More than one codon specifies every amino acid except methionine and tryptophan. AUG in the contest of a ribosome-binding site is the initiation codon and establishes the reading frame that determines the grouping of nucleotides into triplet codons. The code is a non-overlapping code where within a reading frame the first three nucleotides (the start codon) constitute one codon (specify methionine or formyl methionine), the next three (nucleotides 4, 5 and 6 of the reading frame) constitute the second codon (specify the second amino acid of the encoded protein), and the third codon (nucleotides 7, 8, and 9) specify the third amino acid etc etc until a stop codon (nonsense codon) is encountered. At the level of the mRNA protein coding regions are flanked by punctuation codons where the encoded protein sequence begins with a start codon and ends with an in- frame stop codon.

Co linearity of gene and polypeptide Gene expression based on the genetic code produces the co linearity of a gene's coding region and the polypeptide it encodes. There are two major steps in gene expression, transcription is the first and translation is the second.

During transcription RNA polymerase synthesizes a single stranded transcript from the DNA template. This process begins with RNA polymerase binding to the promoter and unwinding the DNA helix to expose bases on the template strand for pairing. Next the polymerase extends the mRNA by adding to its growing or 3-' (prime) end by forming phosphodiester bonds between the 5' carbon of the incoming (next) nucleotide and the 3' carbon of the growing chain. At the end of each transcription unit a transcription termination signal causes RNA polymerase to stop the polymerization process probably by signaling polymerase to dissociate from the DNA template. In prokaryotes the primary transcript is almost always the mRNA that guides polypeptide synthesis. In contrast transcripts in eukaryotes require processing. This processing includes the addition of a cap to the 5' end of the transcript and a polyA tail to the 3' end. Processing in eukaryotes also includes the removal of sequences called introns (derived from intervening) from the primary transcript. Introns are removed and the remaining information, the exons (derived from expressed) are spliced together (joined to form the final processed transcript) to form the mRNA by the intron removing enzyme machine called the spliceosome. For some genes there is more than one way to process the primary transcript into mRNA. This is referred to as alternative splicing (we will explain this later in this course).

Translation is when the cell synthesizes proteins according to instructions in the mRNA. It is called translation because the information coded in the nucleic acid is decoded into proteins, the gene products. tRNAs carry amino acids to the RNA-dependent protein polymerase (ribosome). This is a complex molecular machine composed of many proteins and a few RNA species. Aminoacyl-tRNA synthetases catalyze the bonding of amino acids to the 3' end of their corresponding tRNAs. Each tRNA molecule has an anticodon complementary to the mRNA codon specifying the amino acid it carries. Wobble is the ability of some tRNAs to recognize more than one mRNA codon. Wobble occurs because the 5' nucleotide of the anticodon can pair with more than one nucleotide at the 3' side of the codon. Thus each tRNA has two important sites. One

allows the tRNA to specifically interact with its aminoacyl-tRNA synthetase, which attaches the correct amino acid to the 3í end of the tRNA. The other site is the anticodon that base pairs with specific codons in the mRNA.

Ribosomes have two binding sites (the P site and the A site) for tRNAs and an enzymatic activity that catalyses formation of a peptide bond between amino acids carried by the two tRNAs present in the P and A sites. This enzyme activity is called peptidyl transferase. Translation can be divided into three stages, initiation, elongation, and termination.

Initiation To start translation, part of the ribosome binds to a ribosome-binding site on the mRNA, which includes the initiation codon (usually an ATG). Special initiating tRNAs with anticodons complementary to the AUG carry the amino acid fMet in prokaryotes or Met in eukaryotes to the ribosome's P site. This first amino acid becomes the animo terminus (N-terminus) of the polypeptide.

The initiation codon is AUG. It is distinguished from internal AUGs by a Shine-Dalgarno ribosome binding sequence near the beginning of prokaryotic ORFs and by a cap at the beginning (5' end) of eukaryotic mRNAs.

Elongation The carboxyl group of the amino acid connected to a tRNA at the ribosome's P site becomes transferred and linked by a peptide bond to the amino group of the amino acid carried by the tRNA at the A site. Once the transfer occurs the ribosome moves 3 nucleotides downstream (towards the 3' end of the mRNA). After peptide transfer the tRNA in the P site (the one without an amino acid) is released and the tRNA in the A site is transferred to the P site. This transfer process is associated with the ribosome moving 3 nucleotides towards the 3' end of the mRNA (downstream) and places three new nucleotides (the next codon) in the A site where it can recruit the next aminoacyl tRNA. The elongation process is then repeated. The net result is that the 5' to 3' direction of codons in the mRNA is translated into the collinear amino acid sequence of the peptide (N-terminus to C-terminus).

Termination When the ribosome encounters an in-frame stop codon it ends translation by releasing the peptide from the tRNA in the P site and dissociating from the mRNA.

Coding region or open reading frame The information in mRNA between a gene's start and stop codons.

Mutations When the nucleotide sequence of a gene is changed a mutation occurs. Mutations can be point mutants where one base pair in the DNA is altered to another base pair. Since there are four possible base pairs at any given site, point mutations resulting from the substitution of one base pair for another can lead to three possible changes. Some point mutations change the amino acid sequence of the protein encoded by a gene; these are missense mutations and may or may not affect the encoded protein's function. Some mutations do not alter the amino acid sequence of the protein due to the degeneracy of the genetic code (these are usually silent mutations). Mutations that introduce a stop codon into the region that codes for the protein are called nonsense mutations. Frameshift mutations occur when one or two base pairs are inserted or deleted from the coding region. The protein coding information downstream of the frameshift is read from the wrong reading frame and does not code for the original protein (Note: larger insertions or deletions, as long as they are not multiples of three base pairs, also result in frame shifts).

Gene Structure

Below is the **structure of mRNA in eukaryotes**

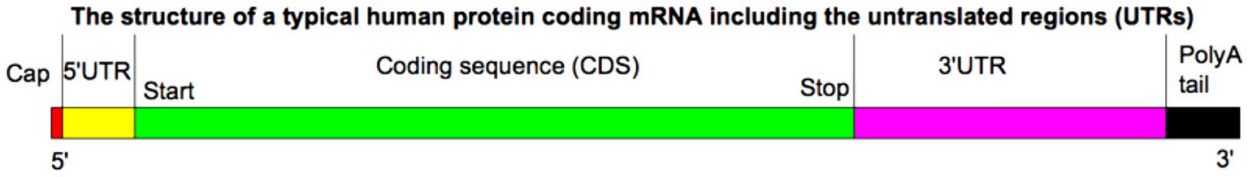
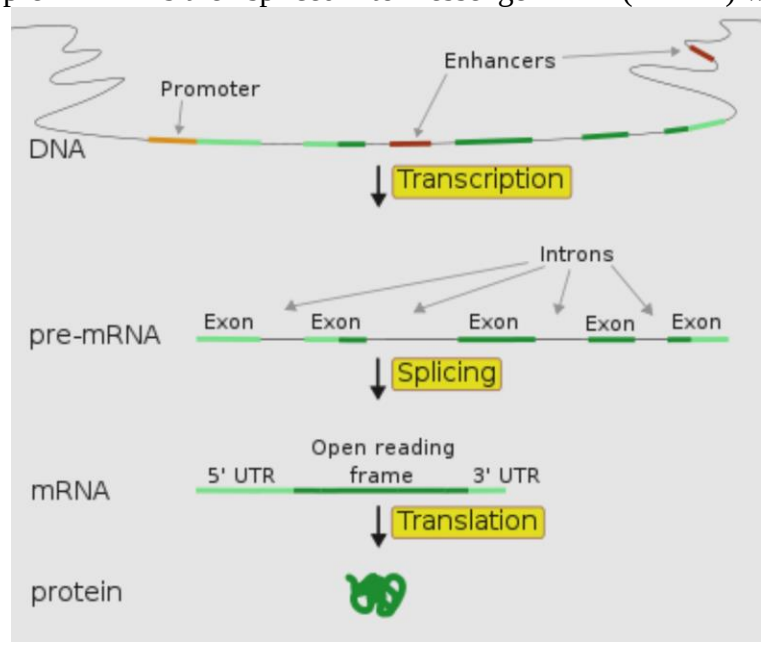


Diagram below of the **eukaryotic protein-coding gen**e. Promoters and enhancers determine what portions of the DNA will be transcribed into the precursor mRNA (pre-mRNA). The pre-mRNA is then spliced into messenger RNA (mRNA) which is later translated into protein



DNA replication (Chapters 20)

DNA replication is **semiconservative** and **semidiscontinuous**. All DNA polymerases require a primer to initiate DNA synthesis. The E. coli chromosome is circular (one DNA molecule per chromosome) and has a single **bi-directional** origin of replication. Eukaryotic chromosomes are linear and have a single DNA molecule per **chromatid**. Replication is initiated at multiple bi-directional replication origins on each chromosome. The region of DNA replicated using a single origin of replication is called a **replicon**. Generally, in prokaryotes there is a single origin per DNA molecule and therefore a single replicon per bacterial chromosome. In striking contrast, eukaryotes have many origins of replication per chromosome and therefore multiple replicons per chromosome.

Enzymology of replication Replication requires that the complementary strands of DNA be melted (separated or unwound). DNA melting associated with chromosome replication is performed by enzymes called helicases (in E coli this activity is encoded by the dnaB gene). The unwound DNA must be kept in a single stranded from until after the replication machinery has completed its replication. To accomplish this unwound DNA complexes with SSBs (single stranded DNA binding proteins). The SSBs aid the helicase by binding cooperatively to the single stranded DNA created by the helicase. Unwinding by the helicase introduces superhelical turns into the DNA ahead of the replication fork. These superhelical turns must be removed. DNA gyrase, a bacterial topoisomerase, is believed to perform this function. A topoisomerase is an enzyme that can remove superhelical turns from DNA (please see the **animation of DNA replication**).

PolIII is the enzyme that replicates E. coli chromosomal DNA There are three DNA polymerases in E. coli, PolI, PolII and PolIII. PolIII carries out the replication of DNA and is the only polymerase that is essential for replication. PolI, PolII and PolIII are all involved in the repair of various types of DNA damage**. PolIII is the enzyme that extends the RNA primers to make both the leading and lagging strands**. It is a remarkable enzyme in that it replicates 1000 nucleotides per second and makes mistakes (incorporates the wrong nucleotide) once in every 1 X 10^{10} bases incorporated (highly processive).

The PolIII core consists of 3 polypeptides, alpha which provides the **DNA polymerase activity**, the epsilon subunit, which has **the 3' to 5' exonuclease activity (required for the proofreading function**), and theta for which a functional role remains to be determined. The polIII holoenzyme is composed of 10 different polypeptides. The polymerase incorporates a wrong nucleotide about once in every 10^5 nucleotides it adds: however, since the enzyme checks (**proofreads**) to make sure that the correct nucleotide has been inserted (monitors the last base pair on the growing DNA to be sure that the structure is consistent with that of properly paired bases) and if a wrong base was inserted removes it by the 3' to 5' exonuclease activity of the epsilon (proofreading) subunit. The net result is that the polymerase gets "two kicks at the can" thereby reducing the error rate to about one mistake in every 10 billion bases polymerized. Even if a mistake remains after polIII has moved on E. coli can detect these replication errors and remove them. This is due to **hemimethylated DNA directed mismatch repair**.

Removing Okazaki fragments In addition to being involved in DNA repair, **PolI** is required to

replace the RNA primers (these are short stretches of about 9 or 10 RNA residues) used to initiate replication at the origin for the leading strand and at the origin and at the beginning of each Okazaki fragment (Okazaki fragments are about 1 kb long) on the lagging strand. To accomplish this **PolI, which is a single polypeptide, has three separate enzymatic activities important for RNA primer replacement, a 5' to 3' exonuclease activity, a DNA-dependent DNA polymerase activity and a 3' to 5' exonuclease activity.** These enzymatic activities are required respectively to remove RNA primers, extend the end of the Okazaki fragment upstream of the primer and to proofread as the newly synthesized DNA is inserted to replace the primer.

Eukaryotic DNA polymerases Mammalian cells contain five different DNA polymerases. We know something about the function of four of them. Two appear to replicate DNA one making the leading strand one the lagging strand. A third replicates the mitochondrial DNA. The fourth is involved in DNA repair.

More detailed analysis of replication The holoenzyme of E. coli DNA PolIII, in addition to being able to synthesize DNA at a rate of 1000 nucleotides per second, is highly processive. **Processivity describes the ability of the polymerase to stick to the template strand during replication. How far the polymerase moves before falling off is a measure of its processivity**. PolIII can apparently replicate the leading strand of the E. coli genome without falling off (remember leading strand replication occurring at each of the two replication forks starts at the origin of replication and ends about 1/2 way around the chromosome).

Replication Complex (multiple enzymes)

1)-Leading strand

- DNA pol -> continuous synthesis on 3' end - initially started with RNA primer

2)-Lagging strand

- must be synthesized in short discontinuous segments
- each segment consists of RNA primer and replicated DNA (Okazaki fragments)
- repeated cycles permit lagging strand to be synthesized
- polymerase removes RNA primer
- ligase seals the discontinuous fragments into a single strand

RESULT: Both chains replicated and continuous

Both strands are replicated simultaneously

Replication Enzymes

DNA Polymerase: Matches the correct nucleotide and then joins adjacent nucleotides together

Primase: Provides and RNA primer to start polymerisation

Ligase: Joins adjacent DNA strands together

Helicase: Unwinds the DNA and melts it

Single Strand Binding Proteins: Keep the DNA single stranded after it has been melted by helicase

Gyrase: A topisomerase that relieves torsional strain in the DNA molecule

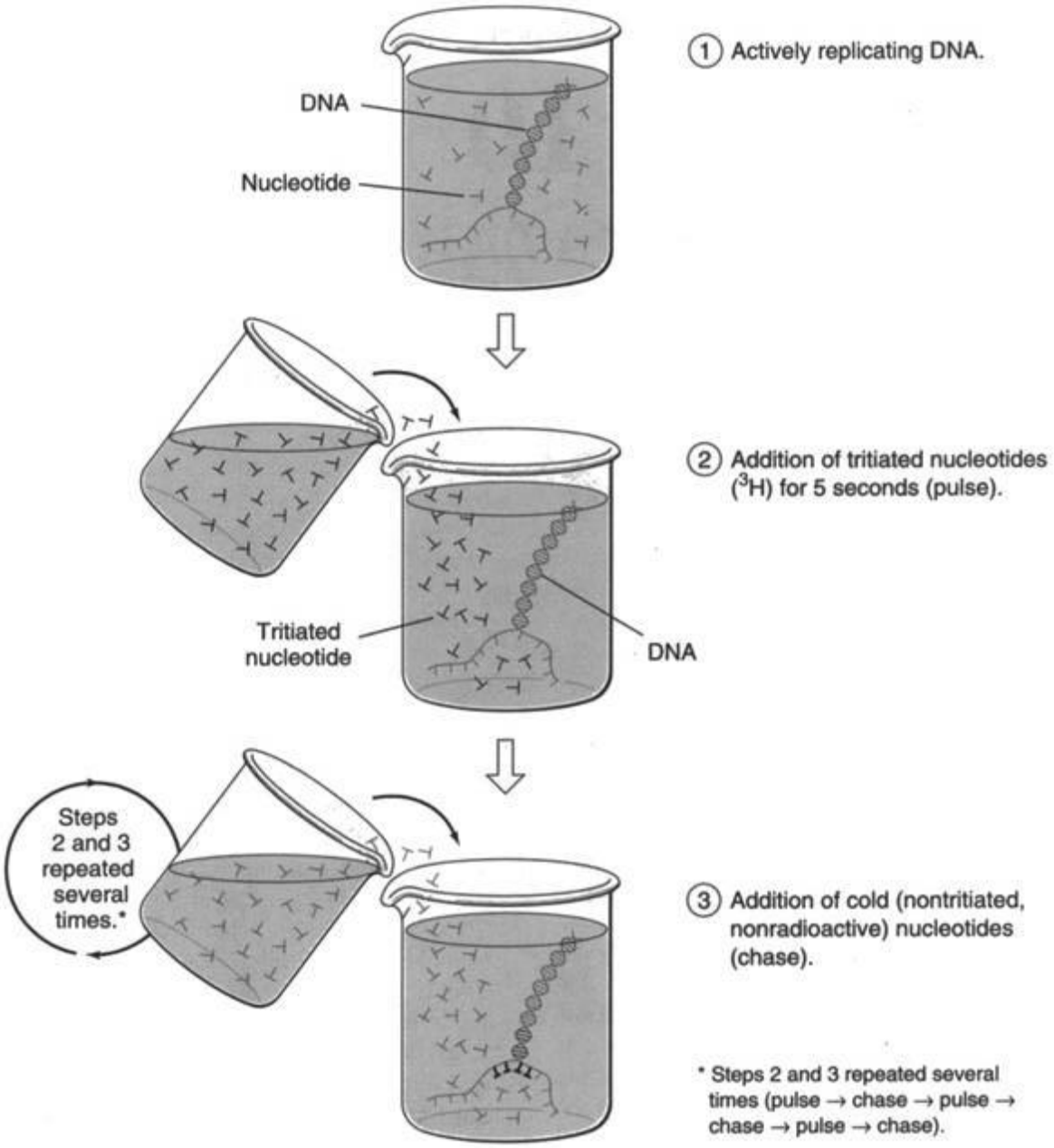
Telomerase: Finishes off the ends of the DNA strand

Prokaryotic replication	**Eukaryotic replication**
semiconservative replication	semiconservative replication
single origin replication (oriC)	multiple origins of replication (ARS)
primer synthesized by primase	primer synthesized by subunits of DNA polymerase α
processing enzyme: DNA polymerase III	processing enzymes: DNA polymerases α and δ
removal of primer: DNA polymerase I	removal of primer: DNA polymerase β
DNA free in cytoplasm as nucleoid	chromatin structure, chromosomes, histones
circular DNA	linear DNA: problem of replication of chromosome ends → telomerase

DNA Polymerase as a Self-Correcting Enzyme

The correct nucleotide has a greater affinity for moving polymerase than the incorrect nucleotide has Exonucleolytic proofreading of DNA polymerase occurs as follows: DNA molecules with mismatched 3’ OH ends are not effective templates because polymerase cannot extend when 3’ OH is not base paired. DNA polymerase has a separate catalytic site that removes unpaired residues at the terminus

OKAZAKI:
DNA SYNTHESIS IS DISCONTINUOUS



(1) Actively replicating DNA.

(2) Addition of tritiated nucleotides (^3H) for 5 seconds (pulse).

(3) Addition of cold (nontritiated, nonradioactive) nucleotides (chase).

* Steps 2 and 3 repeated several times (pulse → chase → pulse → chase → pulse → chase).

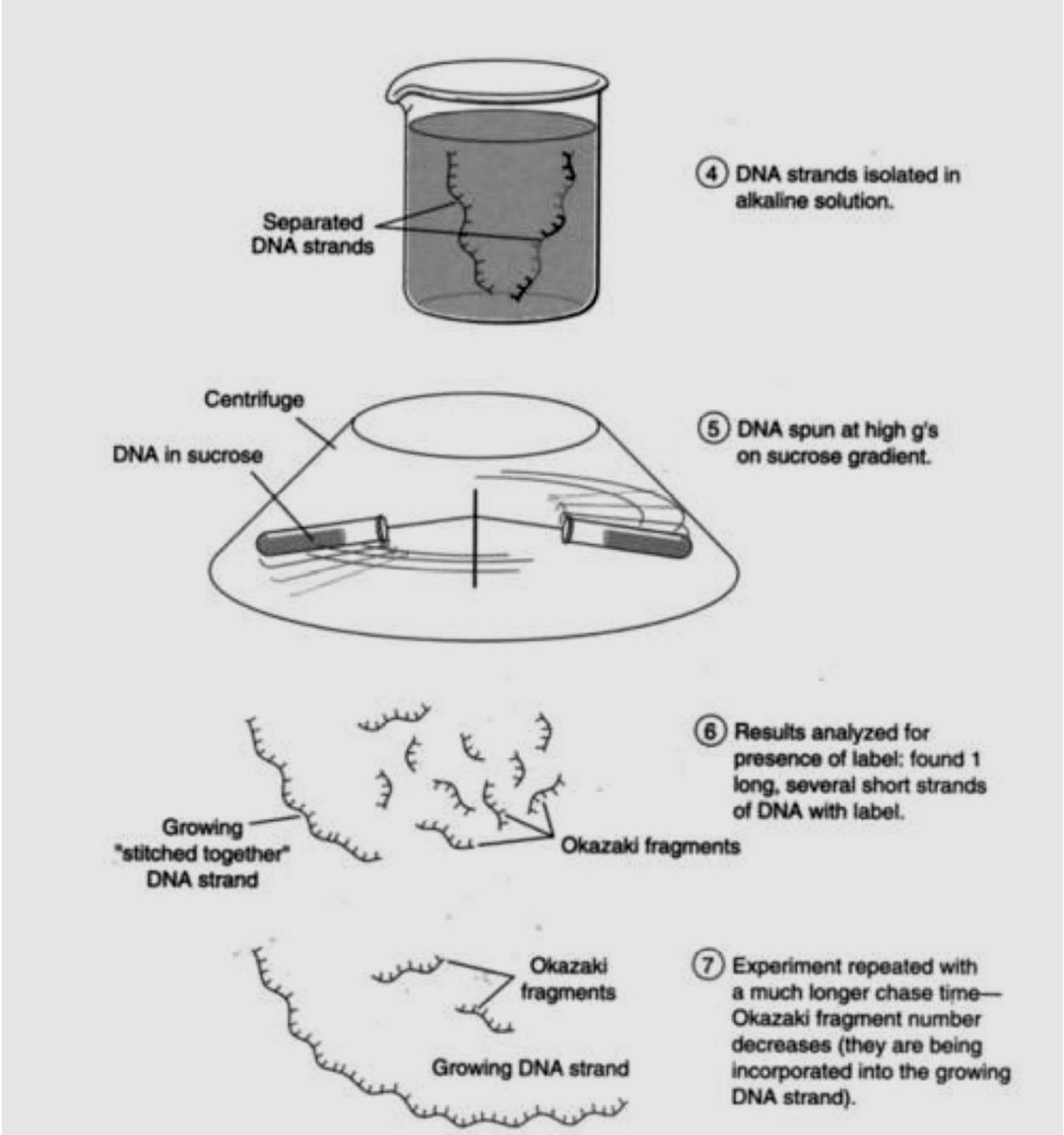
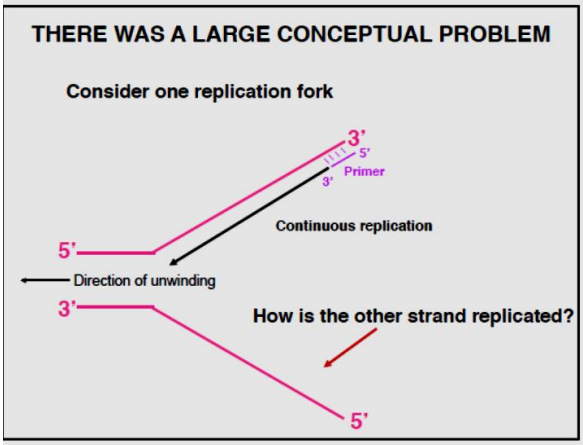
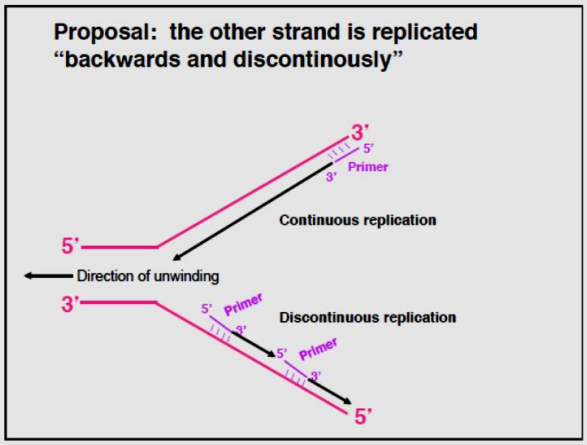


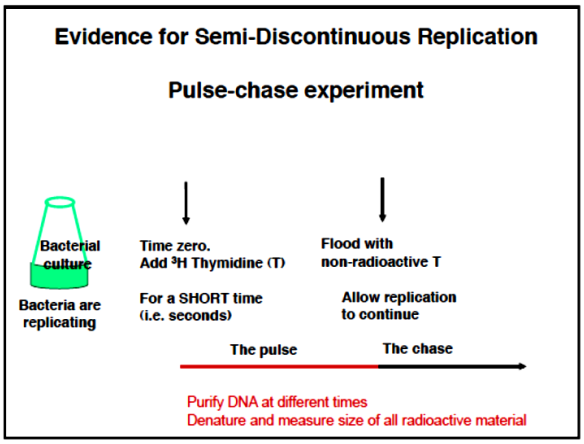
Figure 10.1
Okazaki's experiment.

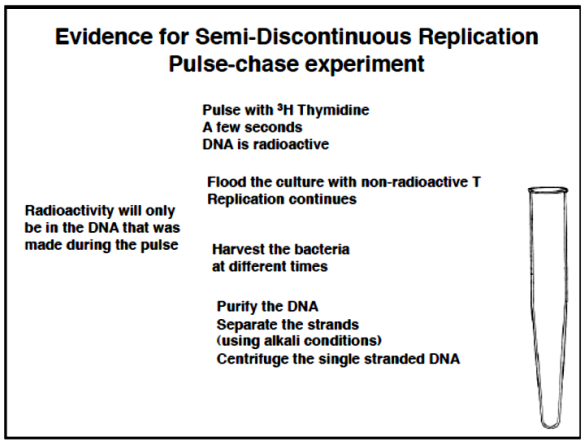
Were the smaller fragments artificially induced breakdown products of normally larger pieces? No: when Okazaki extended the length of the exposure pulse to 30 seconds, a far greater fraction of the total label ended up in long DNA strands. A similar result was obtained if the period of "cold chase" was prolonged prior to isolation of the DNA. **Clearly the fragments**

existed as such only temporarily, and soon became incorporated into the growing DNA strands. As it turns out, normal 5-→3 polymerases are responsible for the synthesis of these Okazaki fragments. Isolation of the fragments and digestion with 3 exonuclease revealed that the label was added at the 3 end of the fragments, as would be expected if the DNA fragments were synthesized by poly-III or another polymerase adding bases at the free 3 –OH end. Finally, the fragments were **joined into DNA strands by a DNA ligase enzyme**, **and mutants that were ligase-negative (lack a functional ligase) failed** to show the pulse-chase assembled into larger fragments.









Evidence for Semi-Discontinuous Replication
Pulse-chase experiment

Centrifuge tube

Contains aqueous solution

Layer the single stranded DNA sample on top

Centrifuge

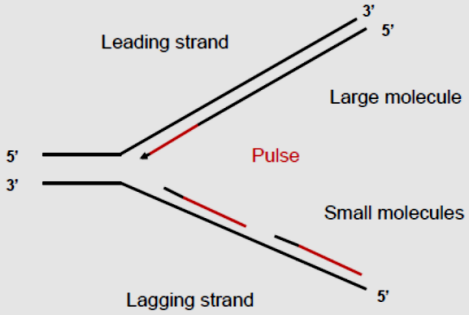
Pierce the tube on the bottom

Collect drops from the tube

Measure the radioactivity in each drop

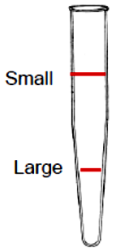
Plot radioactivity per drop

Results of pulse-chase experiment: after the pulse



Evidence for Semi-Discontinuous Replication
Pulse-chase experiment

See *small and large* DNA just after the pulse



DNA purified just after the pulse

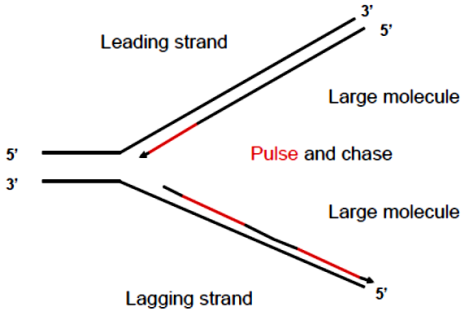
Shows some very large molecules

the leading strand

And some very small ones

the fragments from the lagging strand

Results of pulse-chase experiment: after the chase



Evidence for Semi-Discontinuous Replication
Pulse-chase experiment

See *only large* DNA after a long chase

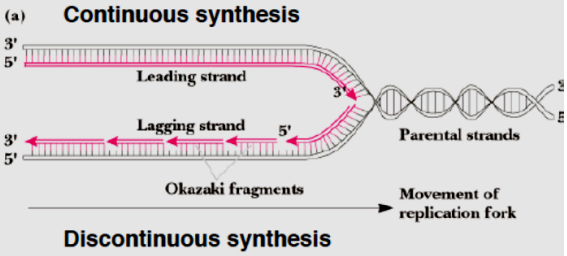
DNA purified just after long chase

Shows only very large molecules

the leading strand

the fragments from the lagging strand have been joined together

DNA replication is semi-discontinuous



Chain elongation is a nucleotidyl-group-transfer reaction

DNA polymerases catalyze the formation of a phosphodiester linkage between incoming dNTP and the growing chain. 3'-hydroxyl group of nascent DNA chain carries out nucleophilic attack on the a-phosphorous of incoming dNTP, forming a nucleoside monophosphate and displacement of pyrophosphate.

- direction of polymerization is 5' to 3'.
- after each addition, enzyme moves by one residue. Process is repeated.
- requires both a template and a primer for synthesis to occur.

Chapter 4: Molecular Cloning Methods

Recombinant DNA technology is the various tools used by researchers to manipulate the very complex genomic DNAs that make up the genomes of organisms. These tools include enzymes for the *in vitro* modification of DNA, methods for the manipulation of DNA and cloning vectors for producing and expressing the products of specific genes.

Restriction endonucleases Type two restriction endonucleases recognize specific sequences in DNA and cut the DNA. These enzymes are useful for cutting DNA molecules and the DNA fragments produced by restriction enzymes are often used for making recombinant DNA molecules in the test tube (in vitro). Many restriction endonucleases make sticky ends (the ends of DNA fragments created by many restriction enzymes have single stranded ends that are complementary to all ends made by the same enzyme). DNAs from different sources that have been prepared by digestion with the same restriction endonuclease can anneal via their complementary stick ends. Once annealed the ends can be covalently linked by the enzyme DNA ligase.

Plasmid cloning vectors pBR322 and the pUC plasmids are useful cloning vectors. Cloning vectors are DNA molecules that have a few specialized features useful for research. For plasmid vectors the useful features include:

Information for replication in bacteria Plasmid vectors replicate autonomously in the bacterial host (usually Escherichia coli). In order to replicate in the bacterial host the vector must include a region of DNA that can be used by the host cell's DNA replication machinery. This region is referred to as an origin of replication.

Unique restriction endonuclease sites Plasmid cloning vectors are used to clone DNA. To facilitate DNA cloning they have unique restriction endonucleases sites (cloning sites) into which foreign DNA can be inserted. Usually cloning a DNA fragment into a plasmid vector involves cutting the vector and the insert DNA by digestion with the same restriction endonucleases. After digestion the inserts and vector are mixed together in the presence of DNA ligase. When the sticky ends of an insert and vector molecule anneal by hydrogen bonding between complementary base

pairs the two molecules for a substrate that can be covalently linked into a single molecule by the action of DNA ligase.

Other vectors discussed We also described cloning vectors derived from lambda (eg Charon 4), cosmids and shuttle vectors for use in yeast and plants.
Shuttle vectors enable DNA to be moved from a bacterial host into another organism.
Clones (many copies of the particular DNA fragment or insert in a cloning vector) are produced when a bacterial transformant harboring a plasmid-insert is replicated by cell division into millions of progeny cells. The resulting plasmid-insert molecules produced are often referred to as DNA clones. Using standard methods the plasmid-insert can be purified away from the other constituents of the bacterial clone (culture of cells derived from the original transformant). The plasmid-insert is purified or separated from the original vector using restriction endonuclease digestion followed by gel electrophoresis to separate the restriction fragments by their size.

The role selectable markers (usually antibiotic resistance genes on plasmids). Why replica plating is used for cloning experiments with pBR322 and screening for white colonies when using pUC plasmids. The physical properties that enable plasmid-cloning vectors to be separated from (purified) other molecules particularly the genomic DNA of the host organism.

Genomic DNA libraries A random collection of genomic DNA fragments from an organism's genomic DNA that have been inserted into a cloning vector is called a genomic DNA library (clone bank). Random fragments can be generated from purified genomic DNA by complete digestion with a type II restriction endonuclease (where the amount of enzyme and length of digestion are sufficient to cut all the sites within the genomic DNA) or by partial digestion (where the amount of restriction endonuclease digestion is only sufficient to cut a portion of the available cut sites). You should know the relative advantages and disadvantages of libraries made with completely digested and partially digested DNAs.
When constructing a genomic DNA library it is important to know how many clones must be constructed. The following formula can be used to estimate how many clones are necessary.

N= ln(1-P)/ln(1-F)
In the above formula N = the number of clones, ln is the natural log, P is the desired probability (i.e. the probability that a specific DNA fragment is preset in the clone bank), and F is the fractional portion of the genome of interest that is present in any given clone. For example, to make genomic DNA library from yeast with the average insert size of 10 kbp (10,000 base pairs long) the fractional portion of the genome in an average clone would be (F = size of inserts/ size of genome or 1×10^3 bp/1.3×10^7 bp).

cDNA libraries Double stranded cDNAs (copy DNAs) are made from single stranded mRNA templates in two steps. The first step involves copying the mRNA template (generating a single strand of complementary DNA) using an mRNA-dependent DNA polymerase (reverse transcriptase). The second step involves generating a second DNA strand using a DNA-dependent polymerase (usually E. coli DNA polymerase I). Once the cDNA copies have been made from the mRNA they are cloned into a vector to make the library.

You should be familiar the relative advantages and disadvantages of cDNA and genomic DNA libraries.

Finding the clone or clones of interest in a cDNA or a genomic DNA library A method has to be devised to find the gene or few genes of interest within the large number of clones in a genomic or cDNA library. There are several approaches that can be used but here we will concentrate on only two.

Approach one If you have access to a sequence that is **homologous** to the gene you are interested (or sequence information about the gene you are interested in) you can make **a probe** for colony hybridization to find the library clones of interest. The homologous sequence can be either a cloned cDNA, a cloned genomic DNA, a PCR product (see below) or an oligonucleotide (a short single stranded DNA synthesized using an automated machine that makes small DNA molecules of the sequence desired). Probes are made by labeling them (making the homologous DNA radioactive). In class we discuss making DNA probes using the Nick Translation method. The radioactive probe is denatured and hybridized to the library of clones. For this the library of clones (individual *E. coli* cells each containing a cloned DNA fragment) is grown into colonies on agar plates. The colonies harboring the library clones are lifted (replicated) onto nitrocellulose filters where they are grown into a second or replicate set of colonies. The replicated colonies on the nitrocellulose must be lysed to release their nucleic acids, some proteins and lipids are removed by treatment with proteases and a detergent and the nucleic acids present in each colony made single stranded by denaturation. The final step in preparation of the filters is to covalently link the nucleic acids from each colony to the nitrocellulose (UV cross-linking can be used). The radioactive probe and the filters with the colony DNA are placed together in an appropriate buffer at a temperature about 20 C below the Tm for the probe when it is hybridized to its complementary sequence. Using an annealing temperature about 20 C below the Tm usually prevents the probe from hybridizing in a nonspecific way with DNA sequences that are not the sequence of interest. The probe should only hybridize to complementary sequences and these should be the sequences of interest. Since the probe DNA was radioactive colonies with the gene of interest can be identified by autoradiography (the radioactive DNA probe will hybridize to colonies with DNA sequences complementary to the probe making them radioactive). To perform the autoradiography following hybridization the filters are washed in a buffer to remove probe that did not hybridize (anneal to complementary DNA). By keeping track of which filters were replicated form which plate and the orientation of the filters it is possible to identify the colonies that have the cloned DNA of interest.

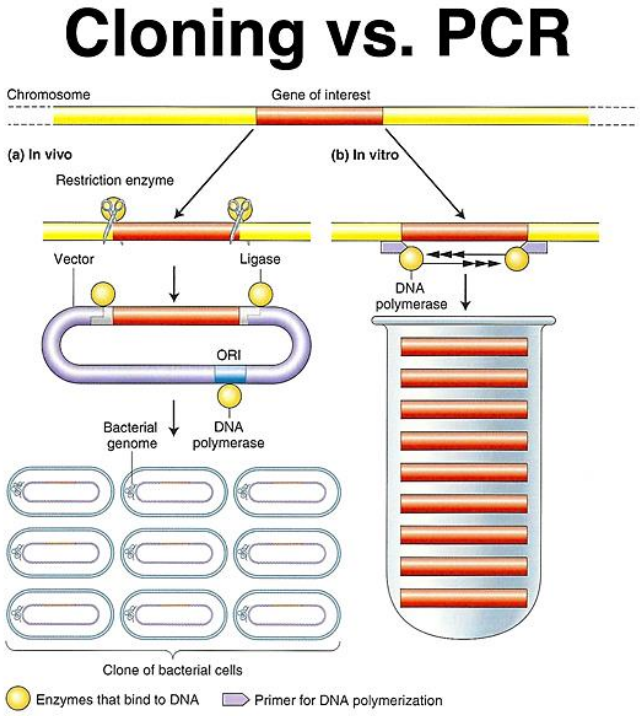
Approach two If the original clone bank was a cDNA library, clones of interest can also be identified by probing with **antibodies** prepared against the protein coded for by the gene of interest. For this approach to be successful it is usually necessary to clone the library inserts adjacent a promoter for expression in E. coli.

Expression vectors One major reason for cloning a gene is so that lots of the gene's protein product can be produced. For this researchers often use expression vectors. To optimize the amount of protein **expressed vectors provide strong promoters and ribosome binding sites just upstream of a cloning site** where cloned DNA coding for the protein of interest can be inserted.

Vectors you should be familiar with Please review the utility and features of the following vectors. pBR322, Charon 4, pBluescript, lambda gt11, His tagging vectors, yeast shuttle vectors YACs and Ti derived plasmids vectors for plants. You should also be familiar with reporter

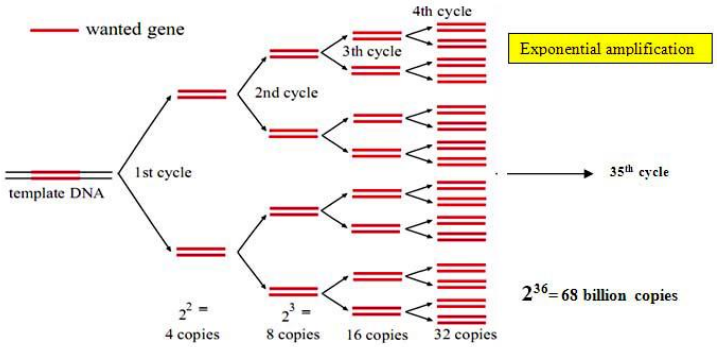
genes (these are discussed below).

PCR The polymerase chain reaction (PCR) was invented by Garry Mullis. PCR amplifies the DNA between two sites that are defined by where the two primers bind to the DNA to be amplified. Each complete cycle of PCR reactions includes, strand denaturation (incubation at a high temp), annealing (incubation at a temp below the Tms for the two primers) and extension (incubation at a temp suitable for the DNA polymerase activity of the heat stable DNA dependent polymerase) doubles the number of copies of the DNA that is flanked by the two primer binding sites on the template DNA.



2 n = number of copies; n=number of cycles -#original copies*2^n

Because both strands are copied during PCR, there is an **exponential** increase of the number of copies of the gene. Suppose there is only one copy of the wanted gene before the cycling starts, after one cycle, **there will be 2 copies, after two cycles, there will be 4 copies, three cycles will result in 8 copies and so on**.



What are some of the uses of PCR in real life?

1. **PCR can amplify a specific portion of DNA** (because of the high selectivity of primer binding to target DNA), **it can be used to isolate a single gene out of the hundreds of thousands of genes in a genome** (just to give you an idea- a gene of 3000 bp (or less) can readily be fished out of a human genome of 3 billion bp). PCR makes isolating individual genes ridiculously easy compared to the methods that were available before PCR.

It's also possible to amplify genes from mRNA after the mRNA has been copied into DNA by using reverse transcriptase.

2. PCR can also be used to introduce mutations into genes, which can then be studied to understand the effects of these mutations on the function of the encoded protein.

3. Since it is possible to amplify large amounts of DNA from tiny traces, it is used by forensic labs to get sufficient DNA from evidence at crime scenes.

4. For the same reason as above, PCR can be used to detect the presence of infectious agents, for example, the AIDS virus (HIV) in blood, long before it can be detected by other means available today.

Genomic DNA libraries

Size of some genomes and chromosomes:

Comparative Sequence Sizes	(Bases)
(yeast chromosome 3)	350 Thousand
Escherichia coli (bacterium) genome	4.6 Million
Largest yeast chromosome now mapped	5.8 Million
Entire yeast genome (completed 5/96)	15 Million
Smallest human chromosome (Y)	50 Million
Largest human chromosome (1)	250 Million
Entire human genome	3 Billion

Fragmentation of genomic DNA for library construction

Restriction endonuclease digestion

- A six-cutter (e.g. Eco RI) will cut on average *every 4.1 Kb*. Complete digestion of human DNA with this type of enzyme will result in approximately 1 x 10^6 unique fragments.
- **What is the probability of finding a clone within a given library?**

The exact probability of having any given DNA sequence in the library can be calculated from the equation

$$N = \ln(1 - P)/\ln(1 - f)$$

P is the desired probability
f is the fractional proportion of the genome in a single recombinant
N is the necessary number of recombinants

For example, how large a library (i.e. how many clones) would you need in order to have a 99% probability of finding a desired sequence represented in a library created by digestion with a 6-cutter?

$$N = \ln(1 - 0.99)/\ln(1 - (4096/3x10^9))$$
$$N = 3.37 \times 10^6 \text{ clones}$$

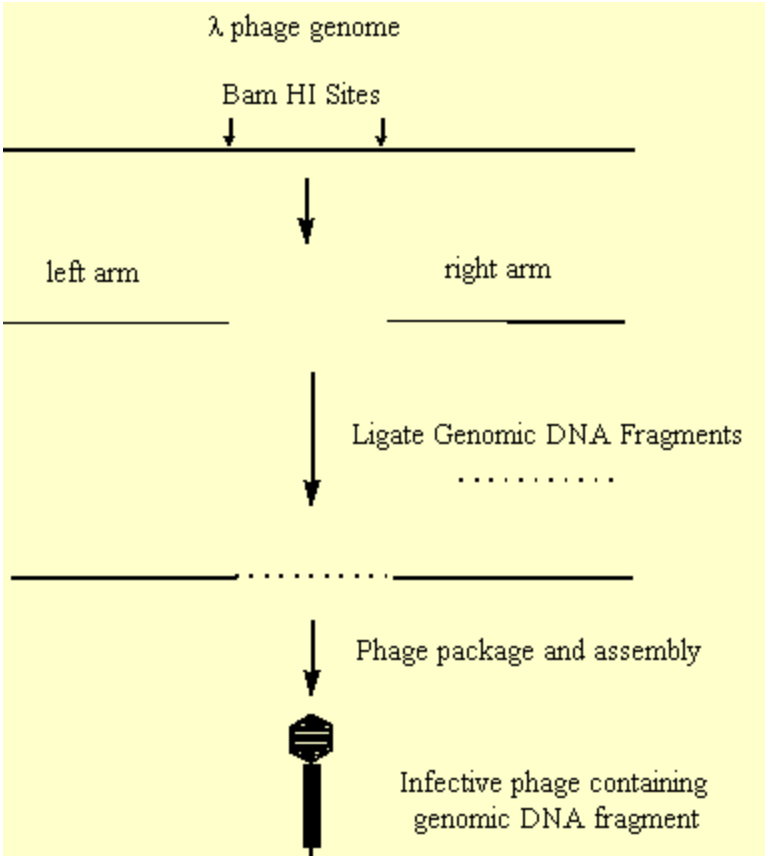
Thus, from this type of analysis we can see that we need a technology which will allow us to achieve the following:

1. *Stable insertion of relatively large DNA fragments into our cloning vector*
2. *High efficiency of insertion and the ability to handle large numbers of clones*

- For example, when plating *E. coli* colonies on a 3" petri plate, the maximum practical density to allow isolation of individual colonies is about 100-200 colonies per plate.
- If we were to try to plate our library of 3.37 x 10^6 in such a way would need about 22,500 plates.
- Not only that, but such large DNA fragments are not well tolerated in typical *E. coli* cloning vectors such as pBR322

Examples of vectors used for genomic libraries:

Bacteriophage lambda vectors are commonly used for construction of genomic libraries

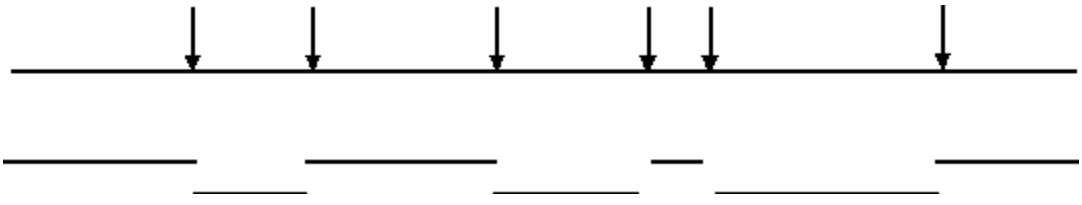


The advantages of this type of system vs plasmids like pBR322 are:

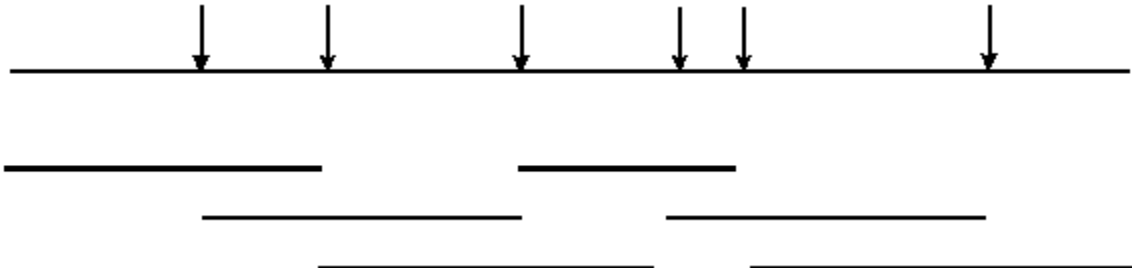
1. *The phage genome is able to package efficiently with DNA inserts as large as 20 Kb*.
2. *Furthermore, the packaged phage are highly infectious and infect E. coli at a much higher efficiency than plasmid transformation methods*.

Incomplete Digestion of Genomic DNA will allow identification of sequence overlaps

Complete digestion with an endonuclease will result in a library containing *no overlapping fragments*:



However, *incomplete digestion* will result in a library containing overlapping fragments:



- Thus, the sequence information obtained from *one clone* will allow the isolation of clones containing *neighboring (overlapping) sequence information*.
- This can allow large contiguous stretches of sequence information to be obtained

Molecular tools for studying genes and gene activity (Chapter 5)

Gel electrophoresis Nucleic acid molecules can be separated according to their size by gel electrophoresis. In gel electrophoresis nucleic acid molecules are size fractionated by subjecting them to an electric field. Because DNA and RNA molecules are negatively charged due to the phosphate groups that link adjacent nucleotides, they migrate towards the positive electrode when in an electric field. The rate at which they move (migrate) in through a porous matrix (the gel) is inversely proportional to their size. Different methods can be used to separate nucleic acids depending on the fragment sizes and whether or not the nucleic acids are denatured (single stranded) or double stranded.

Detecting small amounts of macromolecules, autoradiography and phosphoimaging Detecting the small amounts of substances that molecular biologists often deal with requires the use of labeled tracer compounds. Usually the tracer compound is detected by exposing an X-ray film (autoradiography) although phosphorimaging can also be used.

Southern and Northern analysis Labeled DNA or RNA can be used as a probe to detect nucleic acid molecules of similar sequence using Southern blots (DNA fragments separated by agarose gel electrophoresis and transferred to nitrocellulose membranes) or Northern blots (RNA fragments separated by denaturing agarose gel electrophoresis and transferred to nitrocellulose membranes). With Northern analysis we can estimate the relative levels of a particular RNA species in a population by band intensities (amount of labeled probe that has bound to the Northern blot).

DNA sequencing The ultimate physical map is the complete sequence of the DNA molecule. The chain termination method sequencing method uses dideoxy nucleotides to terminate DNA synthesis. Usually four reaction tubes are employed. All reaction tubes include the template (the DNA to be sequenced), the four deoxynucleotides required for DNA synthesis, a tracer nucleotide (often a radioactive dNTP), **a primer** that binds to the DNA **template** just upstream of the region to be sequenced and a DNA polymerase. In addition a different dideoxynucleotide (dideoxy ATP, CTP, GTP or TTP) is added to each tube. Incorporation of a dideoxy nucleotide causes the growing chain being synthesized by a polymerase molecule to terminate. Therefore, in each tube a series

of discreetly sized fragments are generated. By subjecting the products of each reaction tube to denaturing acrylamide gel electrophoresis the products in the four reaction tubes (each loaded on an adjacent lane) we can fractionate according to their size the various reaction products. Since the products in the four reaction tubes should include all the possible fragments (each differing in size from the next longer or shorter fragment by a single nucleotide) the bands on the autoradiogram can be "read" to tell the DNA sequence downstream of the primer-binding site. Typically about 500 to 1000 bases of sequence can be determined from one set of reaction tubes.

Designer genes and in vitro mutagenesis In vitro mutagenesis enables the precise introduction of mutations into a DNA of interest. Using it it's possible to introduce exact changes into the protein coding or regulatory information of a gene. You should be familiar with the in vitro mutagenesis method (PCR-based mutagenesis) method covered in class (outlined on page 98 of Weaver).

Generating physical maps of DNA The physical map of a DNA molecule, whether it is a cloned genome fragment or a whole genome, is composed of a series of landmarks (something like the towns and cities on a roadmap).
Restriction endonuclease mapping can be used to generate a physical map of a DNA fragment. Gel electrophoresis of fragments generated by digestion of a DNA fragment with restriction endonucleases can be used to generate a map with the location of restriction endonucleases recognition sites on a DNA molecule. Usually more than a single restriction endonuclease is used to produce single digests (cut with one enzyme at a time) and double digests (where two enzymes are used to simultaneously digest the DNA). Southern analysis where fragments generated by one restriction endonuclease (enzyme I) are probed with labeled fragments generated by another enzyme (enzyme II) can reveal important information about how the individual restriction fragments generated by enzymes I and II overlap.

Primer extension Primer extension is a method that hybridizes primers (short synthetic oligonucleotides) to RNA molecules of interest and extends the primers using reverse transcriptase to the 5' end of the RNA template. Once a primer has been extended the resulting DNA molecule is subjected to gel electrophoresis to determine its size and the amount of the extended product. Usually primer extension is performed with a primer that is labeled at its 5' end (made radioactive) by transferring radioactivity form ATP that is labeled (carries P^{32}) on its gamma phosphate using the enzyme polynucleotide kinase.

Based on the size of the extended product the transcription initiation site for the gene under study can be determined. The amount of transcript (relative number of molecules) present in the sample determines how much extension product (number of complementary labeled molecules) was produced. After gel electrophoresis the intensity of bands (on an autoradiogram produced by exposing X-ray film to the gel) that corresponds to the extension products is a measure of the concentration of the transcript of interest in the original RNA sample. This information can therefore be used to estimate the relative amount of transcript present in the sample assayed by primer extension.

Assuming relative proportion of the mRNA species that hybridizes to the primer in a sample isolated from a particular tissue or cell type accurately reflects its relative abundance in

the original cells, primer extension assays measure gene expression (relative amounts of mRNA present in the original cells). Primer extension assays can also measure gene expression in cell free test tube (in vitro) experiments. These in vitro experiments can be used to study the various components that are required for transcription.

Run-off transcription assay The "Run-off transcription assay" is a means of following the efficiency of transcription in an in vitro setting. It can also determine the location of the transcription start site or sites for the gene of interest.

Run-on transcription assay "Nuclear run-on transcription" is a method of assaying gene transcription using isolated nuclei. It therefore approaches being an in vivo method. The method involves isolating nuclei, allowing them to continue gene transcription in the presence of a radioactive RNA precursor (ATP) and then using the transcripts to probe a **dot blot** that includes the genes of interest and appropriate controls.

Methods of reporting gene expression (Reporter genes) The use of two reporter genes, the chloramphenicol acetyl transferase (CAT) gene and the Beta-galactosidase gene (lacZ), will be covered. There are several other reporter genes that have been developed. Reporter genes are useful because they are easy to assay, the cells or organism under investigation does not express the reporter gene and the reporter gene's expression or lack of expression does not affect the system under investigation. Reporter genes are useful because they measure the activity of the promoter region that is fused to them.

Assays for studying the interaction of proteins with DNA These assays are useful because they measure DNA-protein interaction. Two methods of assaying DNA-protein interactions will be discussed here. For both methods the potential DNA target is radioactively labeled and used as a probe.

The filter-binding assay This method relies on the fact that DNA doesn't bind nitrocellulose but protein does. Basically the protein of interest (or a cell extract with the protein of interest) is mixed with a DNA probe that has a potential binding site for the protein of interest. After combining the probe and protein extract the mixture is incubated to allow protein DNA complexes to form. DNA-protein complex formation is measured by assaying how much DNA probe can bind to a nitrocellulose filter. To do this the mixture is filtered through a nitrocellulose filter and the amount of DNA that bound to the filter is measured. The amount of DNA-protein complexes formed (note only DNA with protein bound is retained on the filter since only protein binds nitrocellulose) is estimated by the amount of radioactivity that binds to the filter.

Gel mobility shift assay This method detects interactions between the protein of interest and its DNA target by monitoring electrophoretic mobility of the target DNA (the probe) in the presence of the protein of interest (note the mobility of a naked DNA is greater than the mobility of a DNA that has protein bound to it).

DNase footprinting This method is used to map the DNA sequence that a protein binds to in a sequence specific fashion. The DNA target (probe) is labeled at one end, mixed with the DNA-binding protein of interest, and then the DNA-protein complex is subjected to digestion with the

endonuclease DNase I. The amount of DNase I digestion performed is enough to nick each DNA molecule about once within the region interest. Since DNase I cannot access DNA bound by protein as easily as naked DNA, the protein of interest's binding site can be mapped by comparing the nuclease sensitivity of naked probe DNA and the probe bound with the protein of interest. A footprint or region where the probe DNA from the tube with protein of interests is less sensitive than the control DNA identifies the specific DNA sequence element that binds with the protein of interest.

Determining whether there is a phenotype associated with loss of gene function using knockouts Determining the functional importance of a gene often involves assessing the effect of a **null mutation** of the gene on the organism. Targeted gene replacements and/or targeted gene disruptions can create mutants that cannot express the gene of interest.

The transcription apparatus of prokaryotes (Chapter 6)

Background information on prokaryotic RNA polymerase Bacteria have a single RNA polymerase that transcribes or produces the RNA products (mainly mRNAs tRNAs and rRNAs) by copying the information in genes. There are about 7000 RNA polymerase molecules in an E. coli cell and about 2000 to 5000 are synthesizing RNA at any given time. RNA polymerase holoenzyme (whole enzyme or the active form of the enzyme) is 465 Kd. It is composed of five polypeptide subunits, Beta and beta prime which make up the catalytic center, two alpha subunits which seem important for assembly and for promoter recognition and a sigma subunit (sigma 70 is used for most genes), which is specifically involved with promoter recognition.

Beta, Beta prime, the two alpha subunits plus sigma make up the holoenzyme. The core enzyme (everything except sigma) binds any DNA with a half-life of about 60 minutes but cannot melt DNA nor transcribe (its DNA binding is called loose binding). In contrast holoenzyme binds nonpromoter DNA poorly with a half-life for binding that is less than one second. Holoenzyme binds promoter DNA very well with a half-life of several hours.

Holoenzyme promoter affinity varies depending on the promoter (i.e. promoter strength) for example rRNA gene promoters (very strong promoters) support initiation once every second whereas the lacI gene promoter (a very weak promoter) supports initiation once every 30 minutes (weak promoter).

Polymerase promoter interactions can be described as "closed promoter complexes" (where the DNA is still double stranded) and "open promoter complexes" where a short stretch of DNA is melted by the polymerase. Conversion of a closed complex to an open complex changes polymerase's promoter affinity from loose to tight binding. After open complex formation the binary complex of polymerase and promoter is converted into a ternary complex that includes RNA. Up to 9 bases of RNA can be incorporated without polymerase moving. Release of the short RNA is called abortive initiation. Promoter escape time or clearance time (minimum value is 1 to 2 seconds) defines the maximum frequency of initiation.

About one third of E. coli RNA polymerase exists as holoenzymes because sigma is present in about 1/3 the amounts relative to core enzyme. About half of the polymerases are active in transcription.

Promoter elements Promoter elements are DNA elements that bind RNA polymerase. E. coli

promoters contain two promoter elements centered 10 and 35 base pairs upstream of the transcription start site. These sequences vary between promoters but they resemble to varying degrees (depending upon the gene being examined) TATAAT (for the -10) and TTGACA (for the -35). TATAAT and TTGACA are referred to as **consensus sequences** (the sequence generated when the most commonly present nucleotide at each base pair within an element is used). Determining the consensus for a DNA element requires that several elements be examined. As a general rule the more closely a gene's -10 and -35 sequences resemble these consensus sequences the stronger its promoter.

Some very strong promoters have an additional element that is upstream of the ñ35 element. This element is called a UP element. Its presence can stimulate transcription by as much as 30 times. The seven E. coli genes that express rRNA account for the majority of transcription occurring in rapidly growing cells. These genes have a UP element. Since UP elements are recognized by RNA polymerase, this element is a true promoter element.

Role of sigma Promoter recognition and transcription initiation require sigma but elongation does not require sigma. Indeed sigma is not associated with polymerase during elongation and can be re-used.

Promoter DNA-RNA polymerase interactions There are several distinct steps in the association of RNA polymerase with a promoter and the eventual promoter clearance that leads transcription. These have been termed i) promoter binding (formation of a closed promoter complex), ii) open promoter complex formation (localized melting of about 10 base pairs of DNA, iii) abortive transcription (9 or 10 nucleotides of RNA are repeatedly synthesized without the polymerase moving away from the promoter, and iv) promoter clearance where the polymerase synthesizes more than 10 nucleotides and leaves the promoter region and loses its sigma factor.

Similarity comparisons Comparison of the primary amino acid sequences of different proteins can be used to identify conserved regions and to predict function. Using similarity comparisons between various sigmas researchers were able to focus in on two regions of similarity (called the 4.2 and 2.4 regions respectively). Filter binding studies were able to show that regions 2.4 and 4.2 bind to the -10 and -35 promoter elements respectively. Interestingly sigma can loosen non-specific binding between polymerase and DNA.

UP elements and strong promoters The stimulation of promoter activity by UP elements involves the binding of the UP element by the alpha subunit of RNA polymerase.

Beta and Beta prime subunits The core subunit Beta binds nucleotides at the active site of RNA polymerase. The catalytic region of Beta also has weak affinity (binding activity) for melted or single stranded DNA. The core Beta and Beta prime subunits have a high affinity for DNA downstream of the active site (catalytic region) and Beta prime has a zinc finger that appears to be involved in this strong binding.

Removing excess supercoils from DNA Topoisomerases (enzymes that can introduce transient breaks in DNA so that positive **supercoiling** ahead of the RNA polymerase can be removed) release the strain introduced into the DNA molecule as the RNA polymerase travels along the

DNA template unwinding a short stretch of DNA (melting of about 10 base pairs).

Termination of transcription At the end of a transcription unit RNA polymerase stops transcription and releases the transcript. In E. coli there are two types of transcription termination, rho-independent and rho-dependent.

rho-independent transcription termination (so named because it does not require the protein factor rho) occurs when the polymerase encounters an inverted repeat sequence followed by a series of A residues on the template strand. The inverted repeat can form a stem loop or hairpin structure in the newly synthesized transcript. Hairpins in the newly synthesized RNA cause polymerase to pause. The series of "A" residues in the template DNA weakly interact with the newly synthesized U residues in the transcript. The combination of pausing because the hairpin is formed and weak interactions between the new RNA strand and the template facilitate ternary complex (polymerase-DNA-transcript) dissociation thereby ending the transcription process.

rho-dependent termination requires rho factor. The termination requires an inverted repeat be transcribed from the template strand. The resulting hairpin in the transcript causes RNA polymerase to pause. However the ternary complex does not dissociate spontaneously because the inverted repeat is not followed by a series of A residues in the template strand. Rather dissociation of the ternary complex requires the help of rho factor (thus the rho-dependent nature of termination). To perform its termination function rho has to bind the growing transcript and pursue RNA polymerase as it transcribes. If polymerase pauses then rho can catch up and terminate transcription.

Operons: fine control of prokaryotic transcription (Chapter 7)

Overview Most gene regulation in prokaryotes occurs at the level of blocking (repressor proteins) or enhancing (transcription activator proteins) transcription. Transcription is regulated by the interaction of trans- and cis-acting information.

Trans-acting information (lacI protein and CAP protein are examples) is the diffusible product of a regulatory gene. Trans-acting information therefore codes for transacting factors, usually a protein (could be an RNA). The lac repressor is an example of a trans-acting factor. Because proteins and RNA molecules can diffuse in the cell they can interact with any DNA target sequence as long as the target can be accessed by diffusion, hence there designation as trans-acting (also diffusible).

Cis-acting sequences (a promoter or the lac operator are examples) are acted upon by trans-acting factors. Cis-acting information does not have coding function. It is passive and cannot diffuse to another location in the cell. Cis-acing DNA sequences can only act on contiguous genetic information.

In bacteria (and eukaryotes as we will see later) transcription is regulated by interactions between cis-acting elements and trans-acting factors that occur in the vicinity of the promoter. Although interaction between a promoter (cis-acting) and holoenzyme of RNA polymerase (trans-acting) is fundamental to the expression of all prokaryotic genes, the ability of RNA polymerase to transcribe from a particular promoter can be interfered with by negative acting trans-acting factors (usually repressor proteins) or enhanced by positive acting factors (activator proteins). Repressor proteins prevent RNA polymerase from transcribing genes in two ways. They can block promoters (prevent polymerase binding) by binding sites that overlap promoter

sites (-10 and -35 regions). Repressor proteins can also block the path of RNA polymerase thereby preventing it from proceeding from a promoter through to the genes transcribed from that promoter.

Repressor proteins work by binding to operator sites (short sequences of DNA) found immediately upstream, overlapping or just downstream of the promoter region (-10 -35) for the gene whose transcription the repressor protein regulates.

Repressor proteins come in two types. One type is active as it is synthesized (repressor) and must be inactivated if the gene or genes whose transcription it controls is to be expressed. Repressor inactivation is usually caused by inducers small metabolites that interact with repressor proteins and alter their affinity for operator sites (inducer binding reduces a repressors affinity for its binding site). The other type of repressor is made in an inactive form (aporepressor). Its affinity for its target sequence is too low without the aid of another small metabolite (the corepressor).

How do small metabolites alter DNA binding affinity of a repressor protein (increase site-specific affinity as is the case for a corepressor or decrease it as is the case for and inducer)? Often repressor proteins have two distinct domains. One domain binds a small metabolite the other binds the operator sequence. Binding of the small metabolite to its site causes the repressor to undergo a conformational change (allosteric reaction) that alters the structure (shape) of the DNA binding domain. In the case of inducible genes the small metabolite inactivates the DNA binding domain (reduces its ability to bind to its DNA target). For repressors that are made in an inactive form (the inactive form is called the aporepressor) corepressor binding changes the DNA binding domain's structure so that it can bind its target sequence recognize (bind more effectively to its DNA target).

Positive acting transcription factors enhance transcription. Some promoters are not good binding sites for RNA polymerase. These promoters do not match closely the promoter consensus for RNA polymerase binding. Polymerase binding to these poor promoters requires the help of another protein (such proteins enhance transcription). Positive acting proteins can be also be activated or inactivated by interacting with small metabolites.

Operons In bacteria genes with similar functions are often clustered together and transcribed from a single promoter producing a polycistronic mRNA. A set of genes transcribed from a single promoter is called an operon. Because they are transcribed from the same promoter, the genes of an operon are coordinately regulated. The lactose operon, composed of the z, y and a genes, is expressed as a polycistronic mRNA. These three genes are coordinately regulated at the level of their transcription. lac operon transcription is subject to both positive and negative control.

Negative control of the lac operon Transcription of the lac operon is prevented by the lac repressor protein (lacI), a protein that interferes with transcription of the operon by RNA polymerase. The lac operon's repressor is active (capable of binding its operator site) as translated. For the operon to be expressed the repressor (lacI) must be inactivated. This occurs via a conformational change induced by binding a small metabolite the inducer (allolactose).

Positive control of lac operon expression Positive control of lac operon transcription is exerted by CAP, a protein that enhances transcription. CAP as translated is inactive (is not capable of activating transcription. In order for it to enhance transcription it has to be activated by forming a

complex with a small metabolite cAMP. In its active form it binds just upstream of the lac promoter and helps recruit RNA polymerase to the lac operon's "poor promoter".
Why regulate genes for utilization of carbon sources? E. coli can use a large variety of carbon sources for its energy and carbon requirements. E. coli's preferred carbon source is glucose. The preferred status of glucose probably arose because; it feeds directly into glycolysis, it is widely available in the environment and it is abundant. Lactose, galactose and arabinose, like many alternative carbon sources that E. coli can use, do not feed directly into glycolysis, in fact many alternative carbon sources have to be converted into glucose or an intermediate of glycolysis before they can be used as energy and carbon sources. E. coli discriminates between carbon sources and only uses alternate carbon sources if glucose is not available.

Why bother to discriminate between carbon sources? The reason is probably related to the fact that selective pressure has forced E. coli to be very efficient. Only very efficient microbes have survived to pass on their genes through the billions of years that led to the evolution of E. coli as we know it today. To be able to compete with other organisms and pass on its genes E. coli must be efficient for example it cannot waste energy expressing gene products that it does not require.

Early on researchers learned that the lac operon was only expressed when lactose was present and glucose was not present. Research directed at understanding how E. coli does this not only lead to an in-depth understanding of how the lactose operon is regulated but it has provided mechanistic models that apply not only to E. coli gene regulation but gene regulation in all organisms.

Multiple elements control expression of the lac operon Three cis-acting DNA elements, the promoter, the operator and the CAP binding site control lac operon expression. These three elements are the binding sites for RNA polymerase, the lac repressor protein (i) and the catabolite activator protein (CAP).

The lac operon is subject to negative control by the lac repressor. The lac repressor regulates lac operon so that it is only expressed when lactose is available. Doing this conserves "resources".

How does lac repressor make expression conditional upon lactose availability? lacI (expressed constitutively) has a DNA binding domain that recognizes and binds with high affinity only one sequence in the E. coli genome, the lac operon's operator. Because the lac operator (O) is located at about position +11 (relative to the start of transcription), repressor binding at O creates a roadblock that prevents RNA polymerase from transcribing the z. y and a genes. Repression can be released by adding lactose to the medium. Lactose is converted into allolactose by the enzyme (Beta-galactosidase the product of the lacZ gene). lac repressor has a domain that can bind allolactose. When the repressor protein (i) binds to allolactose it undergoes a conformational change that dramatically alters its ability to bind to the lac O element. That is, when i protein has bound allolactose its affinity for lac O decreases such that the operator site is no longer occupied. Without i bound to the O site the roadblock is removed and the lac operon can be transcribed. Another way of saying this is "it is induced" when lactose is present (note lactose is converted into the actual "inducer" allolactose by the activity of Beta□galactosidase).

Bottom line, lac repressor exists in an active form that can bind its target site (lac O element) and an inactive form when complexed with allolactose. Expression of the lac operon requires the repressor to be in its inactive form.

The lac operon is also subject to catabolite control. Basically catabolite control is

transcriptional regulation that keeps transcription of genes required for the utilization of alternative carbon sources at a low (basal) level when glucose is available. An alternative carbon source is essentially any carbon source other that glucose. We therefore call glucose the preferred carbon source.

How does catabolite control keep lac operon expression low when glucose and lactose are present? The explanation begins with the following fact. The lac operon -10 and -35 elements do not closely match the consensus sequence for these two elements. RNA polymerase binding to the lac promoter is poor or "lousy" (affinity of RNA polymerase is low). Because of this even if the lac repressor has been converted to its low affinity state by binding inducer (allolactose) the lac operon will be expressed at a very low or basal level. High expression only occurs if under conditions that allow RNA polymerase to be efficiently recruited by the lac promoter. Efficient recruitment requires binding of the catabolite activator protein (CAP) to the CAP element. The CAP binding site is immediately upstream of the lac promoter (about position -61). CAP in its active form binds the CAP element and helps recruit RNA polymerase to the lac promoter. CAP assists recruitment because it interacts with the CTD (carboxyl terminal domain) of the RNA polymerase alpha subunit. CAP-cAMP complex bound at the -61 position increases RNA polymerase's affinity for the lousy lac promoter. Catabolite control occurs because the constitutively expressed CAP protein exists in either an "active form" that can bind its DNA element effectively or an "inactive form" that cannot bind its DNA element effectively. Which form it assumes (active or inactive) depends upon whether or not glucose is present.

CAP as translated is unable to bind CAP sites with the high affinity needed for it to occupy CAP sites. High affinity binding of CAP to CAP-sites requires that CAP change its conformation from a form that does not bind DNA effectively to a form that binds effectively. The form it assumes is controlled by intracellular cAMP concentrations. When not associated with cAMP it binds CAP elements ineffectively, whereas when associated with cAMP it binds CAP sites effectively. cAMP levels are controlled by whether or not glucose is available. When glucose is available cAMP levels are low and when glucose is not available cAMP levels are high. Since the CAP site is occupied only when glucose is not available, the efficient recruitment of RNA polymerase to the lac promoter only occurs when glucose is absent. Also the lac operon can only be transcribed when i protein is not bound to the O site.

NB: Biology 367 students should also know what the following lac alleles are i+, i-, is and i-d. Students should also understand how a super-repressor works and how it is that the -d alleles are dominant negative.

Regulation of the tryptophan operon The E. coli tryptophan operon codes for the enzymes required for biosynthesis of the amino acid tryptophan. Tryptophan is essential for protein translation. If cells are growing in a medium with lots of tryptophan it is not necessary to synthesize tryptophan; however, if the medium does not contains enough tryptophan to support growth the cells must produce it themselves. Reflecting this the E. coli trp operon is dual control. One level of control regulates expression in response to the amount of available tryptophan by regulating access of RNA polymerase to the trp promoter. Regulation is also exerted at the level of transcription attenuation.

Regulating access of RNA polymerase to the trp promoter The tryptophan repressor protein is synthesized in an inactive form called an aporepressor. The aporepressor cannot bind to the

tryptophan operator as synthesized it must be converted (its conformation changed) to its active form (form that binds to the operator) by forming a complex with tryptophan (called the corepressor). The aporepressor must associate with the corepressor (small metabolite) to become an active repressor that can bind its target DNA effectively. Repression of the tryptophan operon requires the formation of a complex (the repressor) that is a combination of the aporepressor (protein product of the trpR gene) and the corepressor (tryptophan). In contrast to the lac operon, which is induced by the presence of a small metabolite (allolactose), the trp operon is repressed by the presence of a small metabolite (tryptophan). This makes sense since E. coli does not want to waste resources making tryptophan, an amino acid for protein synthesis, if tryptophan is already available in its environment (medium). The tryptophan repressor accounts for about a 70-fold regulation of the tryptophan operon.

Regulating expression by transcription termination The trp operon is also regulated at the level of attenuation. Attenuation terminates transcription downstream of the promoter but before the structural genes. Attenuation, premature termination of transcription, is signaled by a rho-independent transcription termination signal at the end of the leader region of the trp operon. The termination signal at the level of the DNA is a palindrome followed by a series of A residues on the template strand. Like other transcription termination signals this DNA sequence can produce a transcript with a stem-loop followed by a series of Us. This structure causes RNA polymerase to stall (pause). However formation of this stem-loop is regulated by the concentration of tryptophan in the cell. Attenuation (the use of the transcription termination signal encoded in the leader region) is inefficient during growth in medium with low levels of tryptophan and much more efficient when lots of tryptophan is available. Attenuation loop formation is conditional. Formation of the stem loop that signals transcription termination is prevented if a ribosome pauses at either of two trp codons present in the leader region gene. Formation of the transcription termination loop is not inhibited when the ribosome does not pause at the two trp codons. When tryptophan levels in the cell are low ribosomes translating the leader gene pause longer at the trp codons (waiting for a charged trp tRNA) than when tryptophan and therefore charged trp tRNA are relatively high.
The trp leader region can for alternative stem loop structures Students should understand how tryptophan levels affect the equilibrium between formation of alternative secondary structures within the trp leader region transcript, and how formation of these alternative structures affect transcription attenuation at the end of the leader region.

Major shifts in prokaryotic transcription (Chapter 8)

Prokaryotic systems can undergo dramatic changes in gene expression. Some of the regulatory mechanisms utilized to control global changes in gene expression are outlined below.

Temporal expression of bacterial virus genes Phage systems discussed in class included SPO1 and lambda. Both SPO1 and lambda regulate gene expression such that some genes are expressed immediately after the phage infects the host cell, some begin being expressed (are turned on) after a few minutes (about 5 min) and others “the late genes” are not turned on until about 10 minutes after infection. SPO1 and lambda use very different methods to regulate the temporal expression of these three classes of genes.

Switching promoters is used by several phage to change global gene expression Promoter usage is the method employed by SPO1 and T7 and T3. The SPO1 early genes are recognized by the host polymerase with its associated host encoded sigma. Because the early genes have promoters with -10 and -35 elements that closely match the B. subtilis -10 -35 consensus the early genes can be transcribed as soon as the phage genome enters the host cell. One SPO1 early gene codes for a sigma factor, gp28. This sigma associates with the core RNA polymerase of B. subtilis and forms a new holoenzyme that recognizes promoters with a different -10 -35 sequence (remember it is the sigma factor that determines the DNA sequence specificity of holoenzyme binding). Once this alternate holoenzyme forms the middle genes can be transcribed. Two middle genes encode peptides (gp 33 and gp 34) that can associate with the core RNA polymerase enzyme thereby generating and alternate form of polymerase that a third type of -10 -35 region. This third promoter sequence is associated with the SPO1 late genes. T7 and T3 also switch from early gene expression to late gene expression by switching promoter usage, the method these T phage use is slightly different. The early genes are again expressed using the host polymerase, however one early genes codes for an RNA polymerase (one subunit does everything) that recognizes the promoter sequence used by late genes.

Bacillus subtilis During sporulation of B. subtilis, which occurs during poor growth conditions, switches promoters to effect major changes in gene expression. The method used is to switch promoter usage is by expressing genes that code for alternate sigma factors at different times during sporulation. The new sigmas enable different classes of genes to be expressed as new sigma factors are expressed during the different stages of sporulation.

Controlling changes in global gene expression in E. coli E. coli responds to heat shock stress by releasing a **sequestered** sigma. During growth at normal temperatures E. coli RNA polymerase uses sigma 70. In response to a heat shock a preexisting but sequestered sigma (H with a mass of 32 Kda) is made available so it can compete with sigma 70 for RNA polymerase. The availability of sigma H enables RNA polymerase to interact with a set of promoters that are used to transcribe the heat shock response genes. These genes code for molecular chaperones to refold denatured (partially unfolded) proteins, and proteases to degrade proteins that cannot be refolded.

Lambda phage as a model for studying gene expression Lambda phage, after infection of E. coli, can either enter a lytic cycle (produce many progeny phage and lyse the cell) or enter a lysogenic state where the phage genome integrates into the bacterial chromosome and is maintained as a single integrated copy that is replicated along with the host cell's chromosome. Under normal growth conditions each lambda phage genome that infects E. coli has a 50:50 chance of entering choosing the lytic or lysogenic developmental pathways. Which pathway is followed depends on whether cro or cI wins the race for the operator sites associated with P_R.

Overview of the lytic and lysogenic cycles Immediately after infection P_L (promoter left) and P_R (promoter right) are used to make two short mRNAs (both promoters are recognized by the holoenzyme of the host RNA polymerase). These two short mRNAs code for the N and cro proteins respectively. These two small transcripts from the N and cro genes are called the immediate early transcripts. The immediate early transcripts terminate immediately after these two genes because the host polymerase encounters rho-dependent transcription termination signals just downstream of the N and cro genes. Once levels of the N protein build up

sufficiently it can bind to "nut" sites (N utilization sites) that are present in the N and cro transcripts. After N binds to these transcripts it works with the help of several cellular proteins. Through direct interaction with polymerases transcribing the cro and N genes, N alters the polymerase so that it ignores the rho-dependent transcription termination sites at the end of the two immediate early transcripts (can you suggest how termination is overridden?). This allows the two delayed early transcripts to be expressed from P_L and P_R. These two transcripts code for the N, cIII, gamma, beta, alpha, xis, int and att genes and the cro, cII, O, P and Q genes respectively. Until this point in the infection the decision between lytic and lysogenic pathways has not been made.

The lytic cycle First the events that lead to lysis (cro wins the race) will be outlined. If cro binds OR_3 it blocks transcription from P_{RM} (Promoter Repressor Maintenance). Preventing transcription from P_{RM} blocks cI synthesis from P_{RM}. As we will see below this allows for the build up of the O, P, and Q proteins. O and P code for proteins involved in the replication of the lambda chromosome. The Q protein plays and important regulatory role. A third promoter $P_{R'}$ is also recognized by the host core polymerase. It is used to express the late genes; however, because a transcription termination signal is immediately downstream of $P_{R'}$ this promoter initially directs the synthesis of a very short 194 bases long transcript. Once Q protein reaches a critical level it acts as an antiterminator (remember N is also an antiterminator). Q's mode of action is different, rather than binding to a site on the transcript it binds to a site on the lambda chromosome called the "qut" site (for Q utilization). The polymerase, while transcribing the short 194 base transcript from $P_{R'}$, passes through this site and picks up the DNA bound Q protein. Once Q associates with the polymerase it signals it to ignore the termination signal used to produce the 194 base transcript. Transcription of the late genes leads to the expression of many proteins that are necessary for making new lambda phage and for cell lysis.

The lysogenic cycle For lysogeny to occur cI has to win the race for the operator sites associated with P_R. cI can be expressed from two promoters P_{RE} (Promoter Repressor Establishment) and P_{RM} (Promoter Repressor Maintenance). Neither of the lambda repressor promoters can be transcribed by host RNA polymerase without the assistance of an additional transcription factor. First cI must be expressed from P_{RE}. Two of the delayed early genes are cII and cIII. These two genes code for a transcription factor (cII) and a protein that slows the destruction of cII by host cell proteases (cIII). cII stimulates binding of host polymerase to P_{RE}. Without cII the host polymerase cannot recognize the P_{RE} -10 -35 because it does not resemble the consensus for -10 -35 sequences recognized by the host polymerase. Apparently cII works somewhat like CAP-cAMP and enhances polymerase binding to a poor -10 -35 sequence. DMS interference experiments indicate that cII binds to the opposite side of the DNA helix in the same region as the host polymerase (-21 to -44). cIII retards the action of host proteases that degrade cII. Thus cIII increase the half-life of cII allowing its levels inside the cell to increase. Also, because the transcript produced from P_{RE} is antisense to the cro transcript, its hybridization with cro transcript interferes with cro translation. cI binds to the three operator sites (1, 2 and 3) adjacent P_R. cI binding to operator sites 1 and 2 is **cooperative**. When bound to these sites cI serves two functions; First it prevents the host polymerase from binding to P_R thereby preventing expression from P_R. Secondly, it enables the host polymerase through protein-protein interactions with cI to recognize the poor -10 -35 sequence of P_{RM} thereby enabling transcription from P_{RM}. Once cI occupies the O_R and O_L sites all lambda gene expression is shut down except for cI expression from P_{RM}.

Lysis or lysogeny What determines whether cI or cro wins the race for OR_1 and OR_2 or OR_3 respectively? This is determined by whether cro protein or cI protein reaches the critical concentration required to occupy its operator sites first. If cro occupies OR_3 expression of cI from P_{RM} is repressed. If cI occupies OR_1 and OR_2 cro and the other early gene are repressed and lysogeny occurs. These operators together with the cro and cI proteins act like a switch turning on or off these two developmental pathways. Note there are three OR regions associated with both P_R and P_L. Also, as the lytic cycle progresses cro reaches concentrations that enable it to bind to all operator sites associated with both P_R and P_L. Once all these sites are occupied the delayed early genes are shut down including cII and cIII the genes required for expression of cI from P_{RE}.

DNA-protein interactions (Chapter 9)

We have been looking at gene expression in prokaryotes. The expression and regulation of genes are mainly dependent upon DNA protein interactions and the regulation of DNA binding activities. The specific DNA protein interactions that we have looked at have included various proteins, RNA polymerase, repressor proteins (cI, cro, lacI, araC and trpR) and transcription activators (cI again, CAP-cAMP and araC again) with their target sequences. All of these proteins have much higher affinities for their target sequences than for non-specific DNA sites.

DNA affinity of regulatory proteins Probably all proteins that have a high affinity for a specific sequence (for example the lac repressor for the lac operator sequence) also posses an affinity, although much lower, for any (random) DNA sequence

Specificity The ratio of a DNA binding protein's affinity for its specific site relative to its affinity for other nonspecific sites defines its specificity.
The specificity for the specific site must therefore be great enough to counter balance the vast excess of nonspecific sites. Remember there is one specific site for the lac repressor protein (the lac operator) and 4.2×10^6 nonspecific sites.

Both the amount of a regulatory protein and the number of specific and nonspecific binding sites are also important There must also be a balance between a DNA binding protein's specificity and its concentration in the cell. It must be able to occupy its target site(s) when in its active form (e.g.: lac repressor as it is synthesized) and not occupy its target when it's in its inactive form (e.g. lac repressor with inducer bound).
How does a DNA binding protein bind DNA? Most DNA binding proteins have DNA binding regions that fit into the major groove of the DNA molecule. Certain amino acids in the DNA binding region of the protein (for example the helix turn helix motif of lambda repressor) make specific contacts with the bases in the recognition site (e.g. the operator sequences bound by lambda repressor). These contacts, which are dependent upon the sequence of the DNA binding site, determine the strength of protein-DNA interactions. The strength of this DNA protein interaction can be reduced by changing the sequence of the DNA target site (Oc mutants of the lac operon are an example) or by changing the amino acid sequence of the DNA recognition domain on the protein (many i^- mutants are of this type).

Bottom line When inducer is not present a molecule of repressor is almost always bound to the lac operator and the remainder of the repressor molecules are bound to non-specific sites. Addition of inducer reduces specificity so that the operator cannot compete for repressor with the excess of nonspecific sites. Repressor tetramers are bound to non-specific sites.

Not only is binding specificity important but looping is also apparently important. For years molecular geneticists knew that the lac repressor functioned as a homotetramer. This seemed odd since the lac operator is a palindrome typical of the binding site for a homodimer. Recently it became clear that there are three operator sites associated with the lac operon, the operator site we discussed in class at +11 bp and two additional binding sites at +412 bp and -82 bp. All three sites are important for normal repression of the lac operon. For example mutating both the upstream (-82 bp) and downstream (+412) sites reduces the repression ratio from 1000 to 18.

Eukaryotic RNA polymerases (Chapter 10)

Overview of polymerases in eukaryotes There are three RNA polymerases in all eukaryotes examined to date. PolI is found in the nucleolus (the organelle in the nucleus that contains the rRNA genes). It is responsible for transcription of the 5.8S, 18S and 28S ribosomal RNAs (rRNAs). PolII is responsible for transcription of the protein coding genes (that is it produces the primary transcripts or heterogeneous nuclear RNAs also called hnRNAs) and most of the small nuclear RNAs (snRNAs). PolIII makes precursors to 5S rRNA, the tRNAs and some other small nuclear RNAs.

The three RNA polymerases can be separated by ion-exchange chromatography, and by their relative sensitivity to alpha-amanitin (polII is the most sensitive, polIII has an intermediate sensitivity and polI is very resistant). Alpha-amanitin, produced by the "deathcap mushroom", acts by intercalating between the base pairs of the DNA being transcribed and inhibits transcription of the three polymerases to varying degrees.

The core enzyme of all three RNA polymerases is a complex protein machine composed of 10 to thirteen different subunits depending on the polymerase and the organism examined. Two of the protein subunits are large with masses greater than 100,000 daltons. The RNA polymerase from yeast are the best characterized. Five of the smaller subunits (subunits less than 50 kdaltons) are common to all three polymerases. Deletion of all but two of the 12-polII-subunits genes of yeast produces a lethal phenotype. Ten of the 12 subunit genes are therefore essential for viability (probably because the proteins they code for are essential for polII function) the other two are required under some conditions but not others (we call such mutants conditional mutants).

Two of the polII subunits (RPB1 the largest subunit and RBP6 one of the smaller subunits) are heavily phosphorylated when isolated form cells grown in the presence of radioactive phosphorus. RPB1 (one copy per polymerase molecule), RPB2 (one copy) and RPB3 (two molecules per polymerase II molecule) are all required for polymerase activity and are homologous to Beta', Beta and alpha the three subunits that make up the core enzyme of prokaryotic RNA polymerase. Interestingly the yeast proteins perform similar functions to their prokaryotic homologs. (i.e. RPB1 is apparently the major site for the interaction of polII and

DNA, RPB2 is the site of nucleotide joining and RNA polymerization, and RPB3 serves an organizational function in RNA polymerase assembly by recruiting other subunits and serving an organizational role for core-enzyme assembly.

The large subunit (RPB1) has a CTD (carboxyterminal domain) that is many copies of a seven amino acid sequence (Tyr-Ser-Pro-Thr-Ser-Pro-Ser). The CTD accounts for about 35,000 daltons of its mass. The phosphorylation of RPB1 occurs at serine and threonine residues on its CTD (the importance of this will be discussed later).

The genes transcribed by polI, polII and polIII are designated class I, II and III genes. Each class of gene has it own set of promoter(s). Class II promoters come in various forms. Some have a TATA element about 30 base pairs upstream of where transcription starts (most but not all Class II promoters in lower eukaryotes have TATA elements while in higher eukaryotic plants and animals it is the highly expressed genes that have TATA elements while other genes, perhaps housekeeping genes, tend not to have TATA elements. Some have an initiator region (a consensus sequence that overlaps the location where transcription starts).

We did not spend much time on PolI and PolIII. All polI promoters in a given organism are essentially identical. The general structure of polI promoters is conserved between organisms. They have a core element that surrounds the transcription start region and about 100 bp upstream a second element the UCE (upstream control element). Although the general structure of polI promoters is conserved between eukaryotes the actual sequence of these two elements is not conserved.

PolIII promoters are often within the region transcribed (polIII genes have internal promoters). The 5S rRNA genes have three elements that are important for promoter activity while the tRNA genes have two distinct elements that are important for promoter activity. Some class III genes, e.g.: U6 snRNA gene, actually have TATA elements upstream of the transcription start region.

Gene specific control elements In addition to the information required for the recruitment of RNA polymerase, class II genes have gene specific control elements. All class II genes have enhancer elements (sometimes referred to as upstream activation sequences or UASs). They are usually upstream (but are sometimes downstream even in introns) of the transcription initiation region and are (generally speaking since there are many exceptions) position and orientation independent. Another class of control element, the silencers, are often position and orientation independent (generally speaking with many exceptions). They are used to repress gene expression and are also gene specific. Often eukaryotic genes have multiple enhancer and silencer elements.

General transcription factors in eukaryotes (Chapter 11)

Eukaryotic RNA polymerases (the core enzymes), although they can bind and transcribe from promoters in vitro by themselves, cannot affect transcription in vivo by themselves. They require the assistance of two types of transcription factor that help recruit them to promoters in vivo. These two kinds of transcription factors are called general transcription factors and gene-specific transcription factors.

General transcription factors The additional factors required for transcription of all class II genes by the polII core are called general transcription factors II (TFIIs). By combining basic biochemical methods of protein purification, DNA footprinting and DNA mobility shift experiments (mainly) the general factors that assist polII recruitment to class II promoters have been characterized to varying degrees.

Six distinct TFIIs have been identified and functionally characterized (IID, IIA, IIB, IIF, IIH and IIE). Their structure and function are briefly outlined here. PolII recruitment to class II promoters having TATA element occurs via the following order of binding. IID (perhaps with help of IIA) then IIB, then a complex consisting of polII and IIF (where IIF seems to act as a shepherd which can form protein-protein interactions with the IID + IIA + IIB complex at the promoter and polII) binds. After polII recruitment TFIIE then IIH are recruited.

IID is critical for promoter binding ID from yeast consists of 9 polypeptides ranging in size from 250kd to 30kd. One of these peptides, TBP (38kd) the TATA binding protein, can bind TATA elements without assistance of any other proteins. TBP is unusual; it binds to the TATA element via the minor groove and bends the DNA. The other IID proteins (proteins in addition to TBP) called TAFIIs (for TBP Associated Proteins for polII) are also important for activity. For example TAFIIs 250 and 110 (where the numbers 250 and 110 indicate their molecular masses) help TFIID bind to promoters that lack TATA elements (and perhaps promoters that have TATA elements) by interacting with gene-specific transcription factors (we will discuss these below). In class we looked at IID- SP1 interactions. Different combinations of TAFIIs are apparently capable of interacting with various gene-specific transcription factors.

Summary of pol II general factors Note the different general transcription factors for polII are found in all eukaryotes. Although their subunit structure may vary slightly between eukaryotes their functional activities seem to be conserved such that TFIIs from one organism can provide function when the general transcription apparatus prepared from another organism is depleted for a given TFII.

IID (about 9 subunits), IIA (2 subunits in yeast and 3 subunits in fruit flies and humans), IIB (one subunit), IIF (two subunits one that ushers polII to the growing complex). Once DABFpolII complex has formed polII is capable of forming an open promoter complex and initiating short abortive transcripts. It cannot however perform promoter clearance. Promoter clearance requires factors IIE (4 subunits 2 of a 34 kd polypeptide and 2 of a 56 kd polypeptide) and IIH (at least three subunits, two helicase activities perhaps for unwinding the DNA, and a kinase activity that phosphorylates the CTD of the large subunit of polII perhaps required for weakening the interactions between polII and IID so that promoter clearance can occur). IIE, IIH, IID, IIA and IIB do not appear to be necessary for elongation once promoter clearance has occurred.

Basal transcription apparatus The polII holoenzyme (also called the **basal transcription apparatus**) therefore consists of over 30 subunits. These can be divided into those required for formation of the core enzyme (about 12) and those general transcription factors required for polII recruitment, open promoter complex formation, initiation, promoter clearance and elongation (about 20 polypeptides).

TBP (TATA binding factor) is part of the growing assembly complex for class I class II and class III promoters. For some of these promoters (mainly class II promoters with a TATA) it

has a role in promoter recognition and binding. For those promoters not containing a TATA TBP is still essential; however, it serves an organizational role for the assembly of polI, polII and polIII general transcription factors SL1, TFIID and TFIIIB respectively. The specificity for binding promoter sequences provided by TBP for TATA containing promoters is determined by the other proteins associated with it, TAFs (TBP associated factors) (i.e. TAFIs, TAFIIs or TAFIIIs) when there is no TATA element.

Transcription activators in eukaryotes (Chapter 12)

Gene specific transcription factors are important The basal transcription apparatus (holoenzyme of eukaryotic RNA polymerase II), although capable of directing basal levels of transcription in vitro "cannot transcribe class II genes effectively in vivo". To effect transcription of class II genes in vivo gene-specific transcription factors (also called transcription activators) are also required. All class II genes require the holoenzyme described above for their transcription. The holoenzyme is not sufficient for transcription and requires the assistance of transcription activators.

Transcription activators an overview There are many different transcription activator proteins in eukaryotic organisms and most have two functional domains or regions, a DNA binding domain and a transcription activation domain.

Most DNA binding regions (modules) fall into one of four classes i) zinc finger modules, homeodomain modules, Beta-barrel modules and bZIP bHLH modules. Similarly most transcription-activation domains can be assigned to one of three classes, acidic domains, glutamine-rich domains and proline rich domains.

The DNA binding regions of transcriptional activator proteins act by forming DNA sequence specific interactions with their recognition sequence. These interactions almost always occur within the major groove of the DNA helix using base pair specific interactions similar to those we discussed for the interactions of prokaryotic DNA-binding proteins and their target DNAs.

The transcription activation domain of a transcription activator stimulates transcription through protein-protein interactions. They interact with proteins that are part of the basal transcription apparatus. For example acidic activators like Gal4p and glutamine-rich and proline-rich transcription activation domains appears to stimulate transcription by interactions with TFIIB. By binding IIB they may help to recruit IIB to the growing basal transcription apparatus. The recruitment of IIB also requires IID. Thus these gene specific activators and IID appear to cooperate in the recruitment of IIB. These transcriptional activators also appear to act at a later state of basal apparatus recruitment by interacting with IIF/polII or IIE.

Transcription activator SP1 worked by binding to the TFIID 110 Kd subunit. By interacting with the 110 kd subunit it could help recruit IID to promoters and in vitro allowed polII to be recruited to promoters lacking a TATA element. The take home lesson is that gene-specific transcription factors are able to stimulate transcription in vivo by helping to recruit (assemble) the basal transcription apparatus. This occurs through interactions between the transcription activation domains and one or more of the many proteins that make up the basal transcription apparatus.

The DNA binding and transcription activation domains of transcription activator proteins

are physically and functional distinct regions. Given this it is possible to make functional transcription activators that are composed of the DNA binding domain of one protein and the activation domain of another protein.

Gene specific transcription activators are used to control gene activity The transcription of many eukaryotic genes is controlled by regulating the availability of gene-specific transcription activators or by the activation and inactivation of gene-specific transcription activators. For example, Gal4p (the gene-specific transcription factor required for expression of the yeast galactose catabolism genes) is present but inactive as a transcription activator when yeast cells are grown in the absence of galactose. The inactive form of Gal4p requires another protein Gal80p. Gal80p and Gal4p form a homo-heterotetramer. Gal80p apparently binds to Gal4p adjacent the transcription activation domain. In the absence of galactose the transcription activation domain is not available to recruit the basal transcription apparatus. Note: Gal4p is bound to the gal utilization gene's UASs at all times regardless of whether galactose is present or not. However when galactose is present the tetrameric Gal4p-Gal80p complex changes conformation so that the activation domain is accessible to help recruit the general transcription apparatus.

Gene-specific transcription activators usually form homodimers or heterodimers In addition to transcription activation and DNA binding domains most transcriptional activators have dimerization domains. These domains enable them to form either homo-dimers (dimers formed by complexing with itself) or heterodimers (dimers formed by complexing with another protein). Homodimers bind two identical or very similar half-sites (often the two half-sites form a palindromic DNA binding site). Heterodimers also bind DNA targets with two half-sites but the two half-sites may not be very similar.

Binding as a dimer dramatically increases target sequence affinity. Increased affinity improves or increases binding specificity thereby enabling the target sequence (its specific site, the enhancer or UAS) to effectively compete with the many nonspecific sites that are present in the organism's genome.

Enhancer/ UAS elements Enhancer sites can work at a distance because the DNA between the enhancer and the TATA can be looped out. We saw examples of looping out when we examined regulation of the maltose and lactose operons. Many genes have multiple enhancer/UAS elements. Multiple UAS elements enable a cell to express a gene in response to the expression or activation of several different transcription activators.

Architectural transcription factors Architectural transcription factors, small proteins that have DNA binding domains but lack transcription activation domains, are important for the expression of some eukaryotic genes. They bind and DNA thereby changing the structure of the DNA protein complex assembled at an enhancer-promoter region. For example, LEF1p and HMG1p (High mobility group proteins) bind DNA in the minor groove.

Like with many aspects of eukaryotic and prokaryotic biology there can be many variations on a theme. Above we discussed gene-specific transcription activators that harbored both transcription activator and DNA binding domains on the same polypeptide. It has now become clear that transcription activators can be formed by the interaction of two distinct proteins. One has DNA binding activity the other transcription activation activity. These two

proteins interact through specific protein-protein interaction domains to form a heterodimer (often a tetramer with two activation subunits and two DNA binding subunits) that binds DNA and activates transcription. This arrangement affords more flexibility than transcription activators that are a single polypeptide in that gene expression can be regulated by controlling the availability or the activity of either the DNA binding or the activation portion of the two component transcription activator. Furthermore, both proteins may interact with additional proteins to form other functional two-component activators.

Chromatin structure and its effects on transcription (Chapter 13)

Each human diploid cell contains about 2 meters of DNA. The DNA in all eukaryotic chromosomes is arranged in a very organized fashion with proteins called histones. DNA protein interactions are essential for organizing the huge amounts of nuclear DNA in a manageable form that facilitates replication, chromosome partitioning at mitosis, gene expression, etc. Chromatin is the highly organized combination of DNA and proteins that make up chromosomes. More than 50% of the mass of chromosomes is protein.

Histones, nucleosomes and chromatin organization There are five histones. H1-like proteins vary markedly between tissues and species. Histone H2A shows moderate variation between tissues and species. H2B also shows moderate variation between tissues and species. Histones H3 and H4 are highly conserved between tissues and species.

Nucleosomes The core particle has about 145 bp of DNA wrapped two times around a core of 2 peptides of each of H2A, H2B, H3 and H4. These histone DNA cores are called nucleosomes. The DNA apparently enters and exits the nucleosome on the same side of the nucleosome. In DNA there are about 200 bp of DNA per nucleosome (145 bp wrapped around the core histones and about 55 bp that link adjacent nucleosomes. By combining DNA and the core histones DNA is condensed about seven-fold.

There is one H1 molecule per nucleosome in chromatin. H1 is believed to be located outside the core nucleosome where the DNA enters and leaves the nucleosome core particle.

Higher order organization of chromatin Chromatin is further organized into a 30 nm fiber or solenoid. Histone H1 appears to play an essential role in formation of the 30 nm fibers, although its exact role remains to be elucidated. There are about 5 nucleosomes per turn of the 30 nm fiber with one turn per 10 nm. The 30 nm fiber structure of chromatin is organized as loops with each loop containing between 30 and 90 kbp of DNA. At their ends each loop is anchored to the central matrix or chromosome scaffold. Anchoring is such that each loop can be maintained in a super coiled state.

Histones, nucleosomes and gene activity Nucleosomes and their associated histones are not only important for the packaging and organization of chromosomal DNA they also play an important role in controlling gene activity.

Core histones can assemble spontaneously on naked DNA. In vitro experiments have shown that when enough core histones are added to form one nucleosome every 200 base pairs transcription is repressed to only 25% of the level obtained with naked DNA. Addition of histone H1 (one molecule per 200 bp of DNA) to the core histones further reduces in vitro transcription

to less than 0.5% of the level observed with naked DNA. Core histones and histone H1 can therefore represses transcription. Although the addition of transcription factors like Gal4p and Sp1 cannot counteract the repression effects of the core histones in these in vitro experiments, they can counteract the effect of histone H1. Their ability to counteract H1 is presumably because they compete with histone H1 for binding sites. Thus, at least in an in vitro setting, transcription activators like Gal4p and Sp1 work in two ways. First, they act as transcription activators (we discussed and described this earlier) where they stimulate expression 8 to 10-fold. They also act as antirepressors. In vitro they stimulated expression 25-fold by preventing repression by histone H1. Transcription activators stimulate transcription, when assays are performed with nucleosomal DNA that includes H1, about 200-fold (8-fold as transcription activators X 25 fold as antirepressors).

What is the nature of the relationship between nucleosomes and gene activity? Gene activity is correlated with changes in chromatin structure (**chromatin modifications and chrmoatin remodeling**). Here we will look at different types of remodeling that have been found associated with gene activation.

First, in vivo active genes (genes that are being transcribed) may have control regions (TATA and their associated enhancer regions) that are hypersensitive (much more sensitive that the average chromatin) to nonspecific endonuclease cleavage with enzymes such as DNaseI. This hypersensitivity for some genes is apparently due to the absence of nucleosomes from the control regions (the experiments we looked at in class involved SV40). Somehow nucleosomes can be removed from some active promoters and their upstream control region.

Secondly, active genes often have nucleosomes on their control regions but their positioning is altered upon gene activation. When the control regions of genes that are actively transcribed are compared with the same control regions when these genes are not actively transcribed the nucleosome arrangement is different. When a gene is actively transcribed the nucleosomes tend to be positioned in a precise way whereas when the gene is repressed they are arranged in a random fashion. This is called nucleosome positioning and transcription activators play a role in promoting the precise positioning of nucleosomes in the control regions of active genes.

Third, histone modifications by acetylations tends to be associated with active genes. The strength of the association between histone cores and DNA is related to the interaction of DNA phosphates (negative charges) and basic amino acid residues (positive charges associated with lysine and arginine) on the tails of histones particularly H3 and H4. Histone acetylases add acetyl groups to amino groups on H3 and H4 lysine side chains (note these amino group side chains provide lysine with its positive charge). Histone acetylation therefore weakens DNA-histone interactions. To date several nuclear proteins (found in the nucleus) that acetylate histones on lysine side chains have been discovered (these proteins are called histone acetylases). p55, Gcn5p, CBP/p300 and TAFII250 are examples of histone acetylases. The latter three all interact with transcriptional activators. Histone deacetylases (enzymes that remove acetyl groups also exist and have been shown to strengthen nucleosome-DNA interactions.

What role does the modification of histone lysine side chains play in gene expression? It is believed that histone acetylation allows for chromatin remodeling a process necessary for gene induction. The weakened nucleosome-DNA interactions resulting from acetylation may enhance gene expression in several ways. First, acetylation may enable gene control regions to be cleared of nucleosomes (the SV40 example) so that the gene-specific transcription factors and the

holoenzyme of polII have access to promoters for gene expression. Secondly, acetylation may weaken histone interactions so that the nucleosomes can to be precisely repositioned in such a way that they no longer interfere with the binding of holoenzyme and transcription activators. Nucleosome acetylations can also facilitate the recruitment of TFIID. Finally acetylation may weaken nucleosome-DNA interactions thereby enabling the transcribing polymerase to move along the transcription unit.

mRNA Splicing (Chapter 14)

Genes are often a series of interspersed exons and introns Most genes in many eukaryotes are interrupted by sequences called **introns**. Genes have been found with anywhere from zero to 60 introns. The location of introns in the DNA can be observed directly in the electron microscope using a method called RNA-looping. RNA-looping can be performed between the primary or processed transcript and the gene's DNA or the gene's cDNA. R-looping can also be performed using double-stranded DNA or single-stranded DNA. As a Biol 367 student you should be able to interpret R-looping experiments done using any combination of the above RNA and DNA types.

mRNA in eukaryotes is produced in stages The pool of primary transcripts produced by the first stage of this process, transcription, is called the heterogeneous nuclear RNAs (hnRNAs). HnRNAs include both the intron and exon regions as found in the DNA copy of the gene.

Splicing of nuclear genes The second stage, intron removal is called splicing. Nuclear introns are removed from the hnRNA by an enzyme machine called the spliceosome. Spliceosomes must recognize three consensus sequences associated with each intron in order to remove the intron from the primary transcript. These three consensus sequences are at the 5' and 3' boundaries of the intron and at an internal site called the branch site. These consensus sites are essential for proper splicing, since mutating them prevents intron removal. Introns are removed via an intermediate that is lariat-shaped.

Small nuclear RNAs (snRNAs) are the agents that recognize the three consensus sequences associated with each intron. These snRNAs exist in the cell coupled with proteins in five complexes called small nuclear ribonuclear proteins (snRNPs pronounced snurps). The individual snRNPs are designated U1, U2, U4, U5 and U6. The snRNA in U1 pairs with the 5' splice site. U2 associates with the branch point consensus. Like U1, U6 also associates with the 5' splice site but it also associates with U2. U5 snRNP associates with the last nucleotide of the exon immediately upstream of the intron and the first nucleotide of the exon immediately downstream of the intron (this apparently aligns the two exons for splicing). Finally U4 base-pairs with U6. The role of this is not clear, although it may be to keep U6 associated in the spliceosome complex until it is needed for splicing. The splicing mechanism and how the different snRNPs operate are detailed in Weaver.

Self-splicing introns Some introns are self splicing, that is they can be removed by a mechanism that is catalyzed by the intron itself. Therefore some RNA molecules have catalytic activity. Self-slicing introns can be divided into two types Group I introns and Group II introns. Group II introns splice via a lariat intermediate in a fashion that is very similar to how nuclear introns are

spliced (i.e. they use an A-branched intermediate). Group I introns are spliced out using a different mechanism, where cleavage at the 5' splice junction is via a free G nucleotide that is not part of the intron. For Group I intron splicing there is therefore no lariat structure formed as in nuclear and self splicing Group II introns where the internal A at the branch point forms three phosphate linkages, one each via its 2' 3' and 5' hydroxyls.

The primary transcript produced by some genes can be subjected to alternative splicing. In alternative splicing under some circumstances the set of introns removed from the primary transcript is different from what it is in another condition. A single gene can therefore code for more than one protein.

Capping and polyadenylation (Chapter 15)

The ends of eukaryotic mRNA molecules are modified posttranscriptionally. A cp is added to the 5'-end and a polyA tail is added to the 3' end. These modifications are essential for mRNA function.

Capping mRNA Caps are made in stages, i) removal of a phosphate from the 5' end of the primary transcript, ii) addition of a GMP residue, which serves as the cap iii) methylation of the N7 of the capping GMP iv) the 2' hydroxyl of the penultimate nucleotide is methylated. Capping occurs before the transcript is 30 nucleotides long.

The cap is important It protects the mRNA from degradation, enables the mRNA to be transported out of the nucleus, enables the mRNA to be translated, and is required for splicing of the pre-mRNA.

Trimming and polyadenylation Most eukaryotic mRNAs have a chain of AMP residues at their 3' ends. This chain (average length 250 nucleotides in mammals) is added posttanscriptionally by a polyA polymerase. Transcription extends beyond the polyA addition site. Therefore primary transcripts must be cleaved at the 3' end before addition of the polyA tail.

The addition of a polyA tail begins with trimming the 3' end of the primary transcript (RNA cleavage). This is achieved by an RNA endonuclease that is recruited by the two consensus signals in the primary transcript. In mammals these signals are AAAUAAA followed about 20nt later by a GU rich motif. The actual sequence of these two motifs varies dramatically between eukaryotes.

Cleavage requires several proteins (4 in mammals) and probably the polyA polymerase. Once these proteins assemble near the 3' end of the transcript using the two consensus signals the transcript is cleaved about 8 nucleotides downstream of the AAAUAAA motif. Next polyA polymerase adds a series of A residues. Until about 10 nucleotides are added this process requires the AAAYUAAA motif. A polyA binding protein is required to aid polyA polymerase once the polyA tail is 10 or more residues long. Once polyA-binding protein has been recruited elongation is independent of the AAAYUAAA motif.

In the cytoplasm mRNA polyA ends are dynamic being shortened by an RNase (polyA nuclease) and rebuilt by polyA polymerase. Once the polyA tail is completely removed the complete mRNA is degraded. The rate of RNase activity at the 3' end and of polyadenylation determines at least partially mRNA half-life. This seems to be controlled by sequences in the 3í

UTR (untranslated region downstream of the stop codon) that somehow interact with the RNase to determine the rate at which the polyA tail is removed.

PolyA tail importance There is evidence from studying the splicing of some primary transcripts that the cap and the polyA tail are important for splicing out the first and last introns respectively. The polyA tail is also important because it increases the mRNA half-life and in some instances increases the translatability.

An example of a transcript below (including the segments):

5'CAP--UTR exon1 + intron 1+ exon2+ intron2+ exon3 + intron3 + exon4 + intron4 + exon5 + intron5 + exon6 UTR--3'polyA tail