

Problem Set 4: Molecular Biology

Due Tuesday, April 28 at 12:00 P.M. in the Bi 1 closet

HOMEWORK INSTRUCTIONS

- 1) Turn in your homework stapled to this cover page.
- 2) Use separate sheets of paper for your answers.
- 3) Write or type your answers neatly.
- 4) Put your name on each page of your answers.
- 5) Box your answers, please, so that the grader can find them.

Points may be deducted if you don't follow these instructions!

ANSWER KEY

Name: _____

Section #: _____

Mail Code: _____

TA Names: _____

Date and Time turned in: _____

Number of pages including this one: _____

AFTER YOU FINISH:

Go to the Bi1 moodle site at <http://www.courses.caltech.edu/> and take the homework survey.

There are 4 questions. The number of parts to each question is listed at the beginning of each; be sure to answer all the parts!

Grade:

Problem 1 _____

Problem 2 _____

Problem 3 _____

Problem 4 _____

TOTAL: _____

Problem 1 – Gene regulatory circuits (15 Points – 4 parts)

Bacteriophage lambda can replicate as a prophage or lytically. In the prophage state, the viral DNA is integrated into the bacterial chromosome and is copied once per cell division. In the lytic state, the viral DNA is released from the chromosome and replicates many times. This viral DNA then produces viral coat proteins that enclose the replicated viral genomes to form many new virus particles, which are released when the bacterial cell bursts.

These two states are controlled by the gene regulatory proteins cI and Cro, which are encoded by the virus. In the prophage state, cI is expressed; in the lytic state, Cro is expressed. In addition to regulating the expression of other genes, cI is a repressor of transcription of the gene that encodes Cro, and Cro is a repressor of the gene that encodes cI (Figure 1).

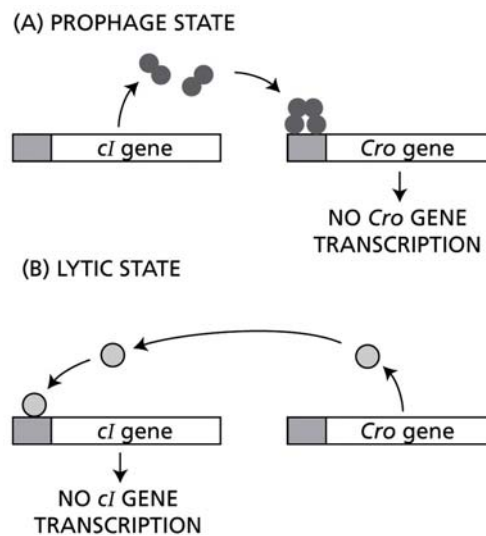


Figure 1: Regulation of bacteriophage lambda replication by cI and Cro. (A) The prophage state. (B) the lytic state.

When bacteria containing a lambda prophage are briefly irradiated with UV light, cI protein is degraded.

A. (3 points) What will happen next? (1-2 sentences)

UV light throws the switch from the prophage to the lytic state. When cI is destroyed, Cro is made and turns off the production of new cI. The virus starts to produce coat proteins, and new virus particles are released.

B. (4 points) Will the change in (A) be reversed when the UV light is switched off? (1-2 sentences)

When the UV light is switched off, the virus remains in the lytic state. Thus, cI and Cro form a gene regulatory switch that, once thrown, is not reversible.

C. (4 points) How is the prophage to lytic switch beneficial to the virus? Hint: Consider the fitnesses of two different viruses: one that is always in the prophage state, and another that is always in the lytic state. **(4 sentences maximum)**

This switch makes sense for the lambda phage. UV light is likely to damage the bacterial DNA, thereby rendering the bacterium an unreliable host for the virus. A prophage will switch to the lytic state, make phage particles, and leave the irradiated cell in search of new, healthier host cells to infect.

Now imagine the two situations shown in Figure 2. In cell I, a transient signal induces the synthesis of protein A, which is a gene activator that turns on many genes including its own. In cell II, a transient signal induces the synthesis of protein R, which is a gene repressor that turns off many genes including its own.

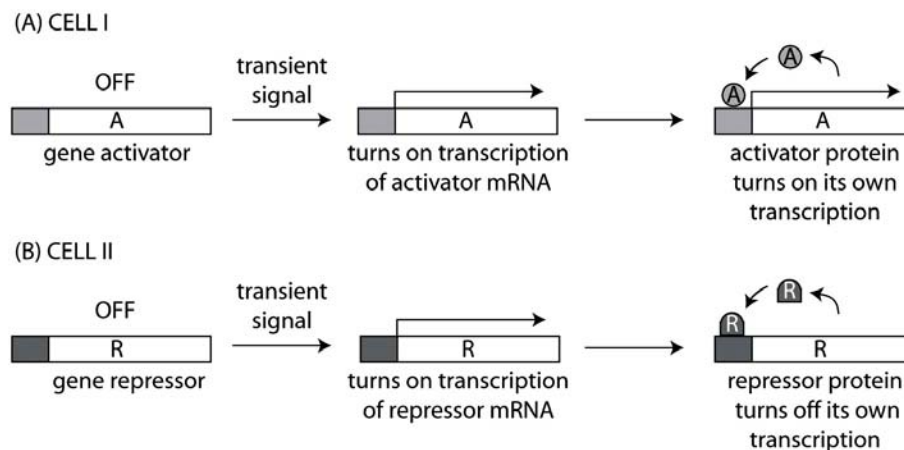


Figure 2: Gene regulatory circuits and cell memory. (A) Induction of synthesis of gene activator A by a transient signal. (B) Induction of synthesis of gene repressor R by a transient signal.

D. (4 points) In which, if either, of these situations will descendants of the original cell ‘remember’ that the progenitor cell had experienced the transient signal? Explain your reasoning.

The induction of a gene activator that stimulates its own synthesis can create a positive feedback loop that can lead to cell memory. The continued self-stimulated synthesis of activator A can, in principle, last for many cell generations, serving as a constant reminder of an event in the distant past. By contrast, the induction of a gene repressor that inhibits its own synthesis creates a negative feedback loop that guarantees a transient response to the transient stimulus. Because repressor R shuts off its own synthesis, the cell will quickly return to the state that existed before the transient signal.

Problem 2 – Polymerase Chain Reaction (PCR) (20 Points – 6 Parts)

Suggested Reading:

- Section 19.2 (3/E and 2/E) from Freeman
- www.idtdna.com/support/technical/TechnicalBulletinPDF/Polymerase_Chain_Reaction.pdf
- www.idtdna.com/support/technical/TechnicalBulletinPDF/A_Basic_PCR_Protocol.pdf
 - The above tutorials are also posted with this problem set.

PCR is a powerful technique that, along with restriction enzymes and other tools, has transformed modern molecular biology through its ability to amplify specific pieces of DNA, even if there is only a small amount of "template" DNA to start with. PCR is used in a wide variety of fields, from basic biological research to forensics to medical diagnosis to cloning DNA fragments from Neanderthals.

Each cycle in the PCR consists of three precisely timed phases that are executed at certain carefully chosen temperatures. Figure 3 is a graph of the imposed reaction temperature as a function of time for any given cycle.

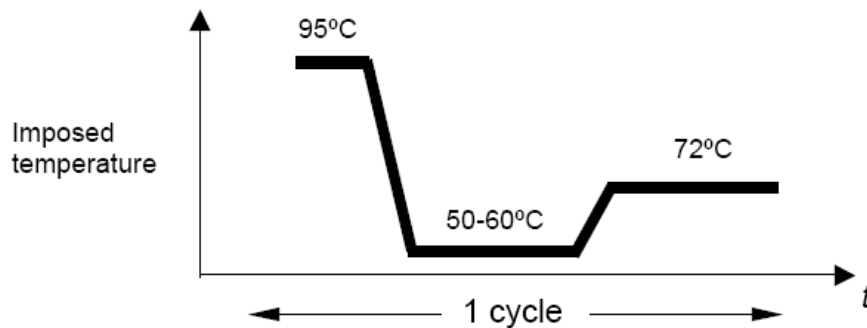
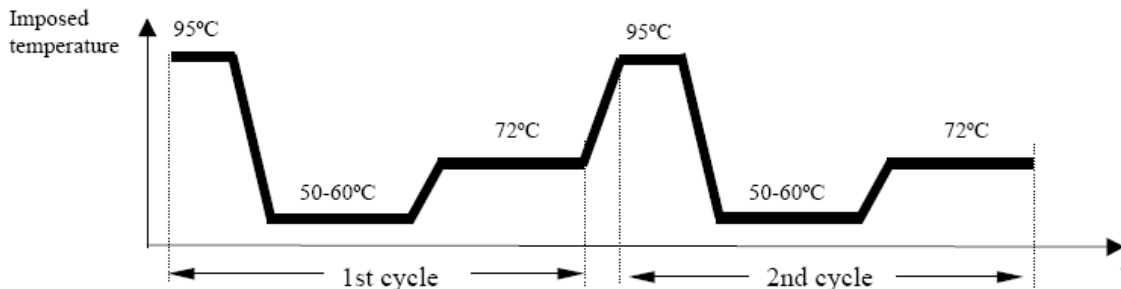


Figure 3: A typical imposed temperature profile during one cycle of a PCR reaction.

A. (2 points) Based on Figure 2, draw a schematic diagram of the temperature profile imposed on the PCR reaction as a function of time for the first two cycles.



B. (4 points) Briefly explain the purpose of each of the three phases and why the specific temperature was chosen for each phase (you may consult your textbook and the references listed at the beginning of this question).

Denaturation (or melting) phase: the reaction mix is heated to ~95°C and double stranded DNA is separated into single strands. No replication takes place at this stage since all DNA is single stranded and the DNA polymerase has no targets to bind to.

Annealing phase – the temperature is reduced to the "annealing temperature" (e.g. 50-60°C) where primers anneal to their specific targets on the single stranded DNAs. The annealing temperature is chosen to be a few degrees below the melting temperature of the primers (i.e. the temperature needed to break the hydrogen bonds of the primers with a complementary targets). Thus as the temperature is reduced, the primers look for good targets to anneal to, and stay annealed to their corresponding targets at the actual annealing temperature. Note that the forward and reverse primers are designed to have similar melting temperatures.

Elongation (or extension) phase - DNA polymerase present in the reaction mix is always looking for free 3' ends present along the DNA to which it can bind to and replicate the single stranded portion of the DNA (DNA polymerase always elongates in the 3' to 5' direction). In the case of PCR, these are the 3' ends of the primers annealed to the single stranded DNA. This process is set at the optimal operational temperature for DNA polymerase, 72°C, although one can also use lower temperatures. Although the elongation temperature is usually above primer melting temperature, the primers do not melt off the DNA since the elongation stabilizes the association of primers to their targets.

Now that you've learned how PCR reactions work, you decide to amplify a segment the HIV-1 gp120 gene (which codes for a surface protein on the HIV envelope) in order to insert it into an expression vector (a plasmid that is used to introduce and express a particular gene). First, you need to design the primers to use for the PCR reaction. Here is a truncated sequence for gp120.

5' TTG TGG GTC ACA GTC TAT TAT GGG GTG CCT GTG TGG AAA GAA GCA ACC ACC.....(middle part of the gene not listed).....CCA TTA GGA CTA GCA CCC ACC AAG GCA AAA AGA AGA GTG GTG CAG AGA GAA AAA AGA 3'

C. (2 points) Which of the following sets of primers should be used to amplify the gp120 sequence?

- | | |
|------------|------------------------------|
| 1. Forward | 5' TTGTGGGTCACAGTCTATTA 3' |
| Reverse | 5' ACCACGTCTCTCTTTTTTCT 3' |
| 2. Forward | 5' AACACCCAGTGTGTCAGATAAT 3' |
| Reverse | 5' ACCACGTCTCTCTTTTTTCT 3' |
| 3. Forward | 5' TTGTGGGTCACAGTCTATTA 3' |
| Reverse | 5' TCTTTTTTCTCTCTGCACCA 3' |
| 4. Forward | 5' AACACCCAGTGTGTCAGATAAT 3' |
| Reverse | 5' TCTTTTTTCTCTCTGCACCA 3' |

D. (4 points) Briefly explain why three of the above options are not suitable to PCR amplify the gene sequence listed above part (C).

- 1: Forward: this primer is correct
Reverse: Wrong polarity
- 2: Forward: this is the complement of the correct sequence
Reverse: Wrong polarity
- 3. Correct
- 4. Forward: this is the complement of the correct sequence
Reverse: this primer is correct

After you have successfully amplified the gp120 coding sequence, you decide that you also want to insert the gene encoding gp41, the portion of the HIV envelope spike that is involved in fusion between the viral and host cell membranes, into an expression vector. You are provided with a DNA sample containing the gp41 gene but you do not know the DNA sequence. You do, however, have the amino acid sequence of the gp41 protein.

gp41 protein sequence:

(N-terminus) Ala Gln Gln His Leu Leu Gln Phe Thr (...middle of sequence not listed).....
Asp Ile Ser Asn Trp Leu Trp Tyr Ile (C-terminus)

You design the following set of primers:

Forward 5' GCG CAA CAG CAT CTG TTG CAA 3'
Reverse 5' TAT ATA CCA CAG CCA GTT CGA 3'

You prepare your PCR reaction, run the gel, and happily see the desired PCR product. However, your friend who is working on the same project, also ran his samples on the same gel as you, but doesn't seem to have a PCR product (no observable band on the gel). Frustrated, he comes to you for help. You compare your methods and realize that your friend has used the following set of primers:

Forward 5' GCT CAG CAA CAC CTT CTT CAG 3'
Reverse 5' AAT GTA CCA AAG CCA ATT ACT 3'

E. (4 points) Give a brief explanation as to why your friend's PCR reaction didn't work.
(Hint: look at the codon table.) (<5 sentences)

The point of this question is to understand that the genetic code is degenerate. Each amino acid has different combinations of codons that can encode it. When only an amino acid sequence is given one does not know which DNA sequence made it. In the first case, the student assumed that the coding sequence is :

GCG CAA CAG CAT CTG TTG CAA...TCG AAC TGG CTG TGG TAT ATA

While the 2nd student assumed it is

GCT CAG CAA CAC CTT CTT CAG ...AGT AAT AGG CTT TGG TAC ATT

The first student got the correct sequence by chance, his primers could have easily been the wrong sequence as well.

F. (4 points) In general, how would you design primers in order to guarantee amplification of a gene if you know the protein sequence but not the DNA sequence? (<5 sentences)

There are several possible answers for this question, if they give a reasonable answer give them credit.

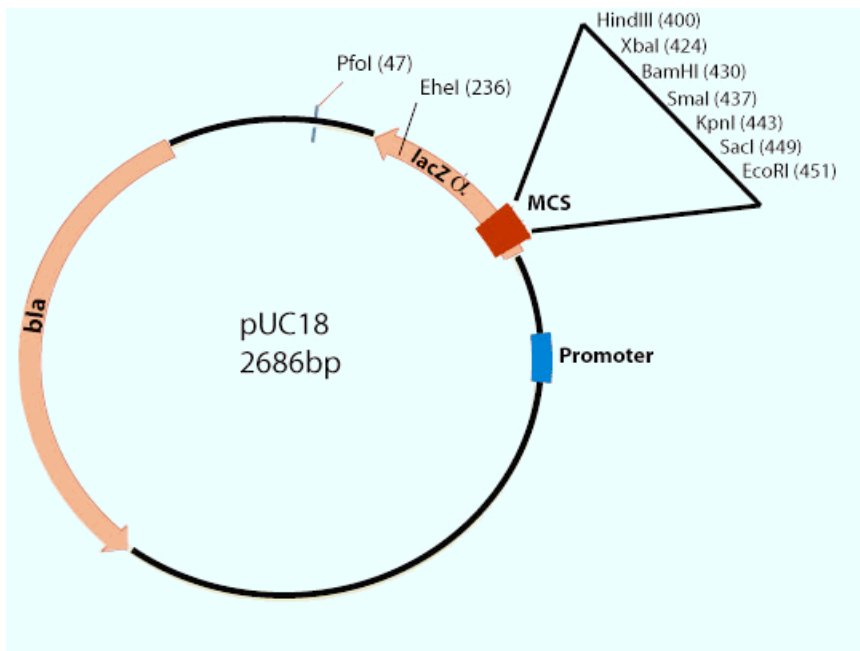
1. Degenerate primers can be used. In degenerate primers, both the forward primer and the reverse primer mix will either have all or most possible combinations of nucleotides. Alternatively, in a more esoteric solution, one can substitute certain nucleotides with universal bases such as inosine, which can base pair with multiple types of nucleotides.
 2. Changing the annealing temperature so that even poorly matched primers bind.
 3. Create primers for a tryptophan rich area cause there is only one codon that codes for it.
 4. Any other logical answers.
-

Problem 3 – Restriction enzymes (25 points – 6 parts)

A common feature of plasmids used in molecular cloning is a short segment of DNA called a “multiple cloning site” (MCS). The MCS allows easy insertion of a new DNA sequence into the plasmid, since an MCS contains many unique restriction sites (i.e., sites that occur only once in the plasmid).

You decide to clone (i.e., insert) your favorite gene into the pUC18 plasmid for expression of the gene product in *E. coli*. To insert your gene, you have designed forward and reverse PCR primers to amplify your gene. Each primer also includes a restriction site (chosen from among the sites that are in the MCS of pUC18). Once you have amplified your gene via PCR, you mix it with a single restriction enzyme (if the restriction sites are the same on both ends) or with two restriction enzymes (if the sites are different) in what is called a “digestion” reaction. You also digest (cut) the pUC18 plasmid with the same restriction enzyme(s), to produce compatible ends for your PCR fragment. You then incubate the digested plasmid and the digested PCR fragment together with an enzyme called DNA ligase. Ligase reseals the breaks after the overhanging ends from the matching restriction digestions anneal so that the PCR fragment becomes covalently inserted into the plasmid. (The ligation reaction is not 100% efficient, so you will end up with some plasmids that do not contain inserts.) The ligase reaction mixture is then transformed (inserted) into *E. coli*, and individual *E. coli* colonies are screened to find those containing a plasmid with an insert.

Below is a (simplified) map of the pUC18 plasmid. Note that the right of the map shows a magnified view of the MCS to show which restriction sites it contains and where they are (numbers in parentheses beside restriction enzyme names refer to the location of the restriction site in the plasmid – i.e., “47” means 47 basepairs from an arbitrary origin). Arrows indicate the direction of transcription of genes.



Important features of pUC18:

- i) β -lactamase gene (bla) to confer resistance to ampicillin
- ii) MCS for inserting genes
- iii) promoter upstream of the MCS for expression of inserted genes
- iv) lacZ α , a lacZ gene fragment that enables “blue/white” screening

Blue/white screening, which is conferred by the lacZ α gene in pUC18, enables easy identification of bacteria carrying plasmids with inserts – blue colonies contain a pUC18 plasmid WITHOUT an insert, and white colonies contain a pUC18 plasmid WITH an insert. Here's how it works:

β -galactosidase (lacZ) is an enzyme that catalyzes the hydrolysis of complex carbohydrates. This enzyme can be split into two peptide fragments: lacZ α (short) and lacZ Ω (long), neither of which is active by itself, but the two fragments can spontaneously assemble to make an active enzyme. The pUC18 plasmid carries coding information for lacZ α . The MCS of pUC18 is embedded near the beginning of the coding region of lacZ α in such a way that the lacZ α reading frame is not disrupted. Thus although extra amino acids are introduced by translation of the MCS, adding a few extra residues near the N-terminus of lacZ α does not affect its structure, its ability to assemble with lacZ Ω , or the enzymatic activity of the assembled β -galactosidase protein. When pUC18 is transformed into an *E. coli* strain carrying the lacZ Ω gene, active β -galactosidase is produced, which can cleave X-gal, a chemical substrate that is included in the agar plate on which the bacteria are grown. X-gal is colorless, but when cleaved by β -galactosidase, it produces a product with an intense blue color. Thus, plasmids with an intact lacZ α fragment will produce blue bacterial colonies when transformed into bacteria grown on an X-gal plate.

If DNA has been inserted into the MCS of pUC18, it almost always destroys the lacZ α gene so that it can no longer produce a lacZ α protein that assembles with lacZ Ω to produce active β -galactosidase. Thus when a pUC18 plasmid containing an insert is transformed into the *E. coli* strain carrying the lacZ Ω gene, no active β -galactosidase is produced, so the resulting colonies are white.

- A. (5 points)** In order for blue/white selection to work, the presence of an MCS lacking an insert must not affect the function of the lacZ α gene. **Aside from being relatively small, how is it that the insertless MCS does not disrupt the lacZ α gene (i.e., does not prevent the gene from encoding a functional protein)?** Your answer should include two required features of the insertless MCS. In order to answer this question, it may be helpful to think about the ways an inserted segment of DNA could be disruptive to the rest of a gene.
- i. MCS in multiples of 3 (in frame with the lacZ(gene))
 - ii. No in-frame stop codons

- B. (2 points)** What is the purpose of the antibiotic resistance gene in pUC18?

So that only bacteria with the plasmid will grow in a media supplemented by antibiotics (a selection for bacteria that took up the plasmid)

- C. (2 points)** Why is it important that restriction sites within the MCS are unique (found only once in the plasmid)? Once you have chosen candidate restriction site(s) to use for inserting your gene, why is it important to check the sequence of

your gene to see if it contains those restriction site(s)?

If the enzyme cuts more than once within the plasmid, you will release a fragment (or fragments). You won't be ligating your insert to a single defined place in the plasmid anymore, so you will get a mixture of products. In some cases, the plasmid will no longer be replicated because you have destroyed the origin or replication or another important feature.

D. (3 points) If you want to express a protein from the inserted gene, why is it advantageous to use two different restriction enzymes, rather than a single restriction enzyme, in the cloning procedure to insert your gene into the MCS?

If you use only one restriction enzyme for cloning, your insert could be ligated in either the forward direction (along transcription of your gene) or in the reverse direction (in which case you will transcribe from the wrong strand, so your protein won't get expressed). By using two restriction enzymes, you guarantee the correct orientation of your inserted gene. Additionally, digesting the plasmid with only one enzyme may cause the plasmid to close back on itself with no insert, the use of two different restriction sites could avoid this (assuming they are not both blunt or leave the same sticky ends).

E. (2 points) You decide to clone your PCR fragment between the EcoRI and HindIII restriction sites found within the MCS of the pUC18 plasmid. Which restriction site would you place at the 5' end of the insert gene? Which site would you place at the 3' end?

5' EcoRI
3' HindIII

F. (11 points total) On the following page is a diagram of an agarose gel used in electrophoresis of DNA fragments produced from restriction digests of the two plasmids in your experiment (pUC18 without your gene, and pUC18 with your gene inserted between the EcoRI and HindIII restriction sites). Use this diagram to answer the following questions; include the approximate sizes of any fragments you draw. Note that the first and last lanes in the gel show a DNA ladder, which is a set of DNA molecules of different (known) lengths that are used as references to estimate the sizes of experimental DNA fragments. Use the ladder to estimate where your restriction products would be found on the gel and to determine the molecular weights of the digestion products in Lanes 7 and 8. For comparative purposes, Lane 9 shows the migration of an uncut pUC18 plasmid that contains an insert. An uncut plasmid does not migrate in proportion to its molecular weight, and it migrates as a mixture of nicked, relaxed, and supercoiled species. Assume the size of your insert gene is 1000 basepairs (bp) and that the pUC18 plasmid is 2686 bp.

i. (1 point) Draw where the positive and negative electrodes should be.
positive electrode at the bottom of the gel

- ii. (1 point) In lane 1, draw the band(s) corresponding to the plasmid (without your gene cloned into it) digested with XbaI and EheI.
two bands~ 188bp and 2498bp
- iii. (1 point) In lane 2, draw the band(s) corresponding to the plasmid (without your gene cloned into it) digested with EcoRI and HindIII.
two bands ~2641bp and 45bp (the digestion removes ~45 bases)
- iv. (1 point) In lane 3, draw the band(s) corresponding to your gene. Assume that your gene is the major product of a PCR amplification and that you have already purified it from minor products and from the starting components of a PCR reaction.
one band 1000bp
- v. (1 point) In lane 4, draw the band(s) corresponding to digestion of the plasmid with SacI. Assume that your gene has been cloned into the plasmid. Also assume that the restriction site for SacI is not found within your gene.
Should match the bands in column 9 of undigested plasmid with insert since SacI was cut out of the plasmid when the gene was cloned into it.
- vi. (1 point) In lane 5, draw the band(s) corresponding to digestion of the plasmid with your gene cloned into it, digested with EcoRI and HindIII.
two bands, 2641bp and 1000bp
- vii. (1 point) In lane 6, draw the band(s) corresponding to digestion of the plasmid with your gene cloned into it, digested with EcoRI and EheI. Assume that the restriction site for EheI is not found within your gene.
two bands ~1164 (gene+lacZ α), ~2477bp (plasmid without lacZ α)
- viii. (2 points) Lane 7 shows the digestion of your plasmid with your gene cloned into it digested with KpnI (refer to MCS on the plasmid map). **Explain the result seen in the gel.**
KpnI is found within the MCS and was cut out for cloning our gene into it.
Therefore there is no KpnI site in the plasmid. Since there is only one band, this means your gene had a KpnI site. Or a mutation occurred in the plasmid and it has another KpnI site.
- ix. (2 points) Lane 8 shows the digestion of your plasmid with your gene cloned into it digested with PfoI. **Explain the result seen in the gel.**
PfoI has a site at 850bp from 5' end within your gene. This leaves us with two fragments: ~506bp and 3135bp

**Problem 4 – Chromatin immunoprecipitation and the control of gene expression
(40 points – 11 parts)**

Gene expression in cells needs to be tightly regulated so as to allow different cells to perform specific functions at the correct times. This is especially important in embryonic development, when undifferentiated cells undergo a differentiation (or fate-determining) step in which they commit to become a certain cell type. For example, during the differentiation of progenitor cells into muscle cells, the number of mitochondria in the cells is increased to satisfy the high energy requirements of muscle tissue. This requires an increase in the number of mitochondrial proteins like cytochrome C, many of which are encoded in the nuclear DNA (as opposed to the mitochondrial DNA).

A. (2 points) What is cytochrome C's function in mitochondria (*i.e.*, how does it help meet the cell's energy requirements)?

Cytochrome C is a mitochondrial membrane protein that resides along the electron transport chain and is involved in the production of ATP. According to Figure 9.24 of the textbook, it shuttles electrons from Complex III to Complex IV, which are proton pumps that use the energy of redox reactions to establish a proton gradient across the mitochondrial membrane. This proton gradient drives ATP synthase to produce ATP.

Researchers wished to understand by what mechanisms the cell increases (“upregulates”) the number of cytochrome C proteins. One possibility is that the cell increases the transcription rate of the cytochrome C gene (called *CytC*), a process referred to as “transcriptional regulation.” Transcriptional regulation can be controlled by transcription factors, which are a class of DNA-binding proteins that can activate or repress the transcription of nearby genes. Each transcription factor has a preferred DNA sequence to which it binds. However, for reasons that may be related to DNA packaging and accessibility, not all such DNA sequences that exist are actually bound by a transcription factor.

Researchers identified a predicted binding site for CREB, an activating transcription factor, upstream of the *CytC* gene. To determine if CREB actually binds to this site, they employed a technique called **Chromatin Immunoprecipitation (ChIP)**. Essentially, cells are fixed, homogenized, and their DNA is sheared into 1-kilobase-long fragments. An antibody that binds to the transcription factor of interest is used to selectively pull out the desired transcription factor:DNA complexes (a procedure called “immunoprecipitation”). The DNA is then released from the transcription factor, allowing its sequence to be determined. [If you have trouble understanding this paragraph, please consult the more detailed explanation of the ChIP procedure at the end of this question, which also includes terminology definitions and a diagram of the steps in the ChIP procedure.]

B. (4 points) How might you determine whether the *CytC* gene is contained within the immunoprecipitated DNA?

Either answer is acceptable:

- a) Use *CytC*-specific primers and PCR to determine whether *CytC* DNA is found in the immunoprecipitated fraction.
- b) Sequence the whole pool of immunoprecipitated DNA and determine if the *CytC* gene is within the pool by aligning the sequences to the genome.

C. (6 points) Describe a good negative control for this experiment. What experimental uncertainty would this control eliminate? Hint: look at the diagram and description of a ChIP procedure at the end of this question.

Either answer is acceptable:

- a) A good negative control is to perform the ChIP procedure except with the antibody omitted. You should not get a band/signal for *CytC* in this case. This control rules out carryover of unbound genomic DNA during the immunoprecipitation procedure.
- b) Another acceptable negative control is to perform the ChIP procedure on a cell sample that has a knockout of the CREB protein, if available. This control will rule out spurious binding of the anti-CREB antibody used in the immunoprecipitation.

In their actual experiment, the researchers performed ChIP on three separate cell populations: undifferentiated cells, differentiating cells, and fully differentiated muscle cells. For their ChIP, they used two different antibodies that specifically recognized one of two similar, but chemically distinguishable, forms of CREB: CREB1 Δ or CREB1 α . Thus the anti-CREB1 Δ antibody bound to CREB1 Δ but not to CREB1 α , and the anti-CREB1 α antibody bound to CREB1 α but not to CREB1 Δ . (CREB1 Δ or CREB1 α are splice variants of the CREB protein, meaning that they are derived from the same pre-mRNA that is subsequently subjected to alternative splicing.) The researchers then determined the amount of *CytC* DNA found in their immunoprecipitated samples:

	Undifferentiated	Differentiating	Differentiated
Anti-CREB1 Δ	+++	++	+
Anti-CREB1 α	0	+++	+++

0 = no DNA found; +, ++, +++ = gradations of DNA found, with +++ being the most

D. (3 point) Which forms of CREB bind to the *CytC* promoter in each of the different cell populations?

In undifferentiated cells, CREB1 Δ binds to the *CytC* promoter but not CREB1 α . In differentiating cells, both CREB1 Δ and CREB1 α bind to the *CytC* promoter. In differentiated cells, most of the binding to the *CytC* promoter is performed by CREB1 α , with only a small contribution by CREB1 Δ .

E. (2 points) Which form of CREB would you say is primarily responsible for upregulating (increasing) CytC expression during muscle cell differentiation?

CREB1 α (note to graders: no explanation is requested in the question, so do not mark off if none is provided)

F. (4 points) Propose a way in which a cell can favor the use of one protein splice variant over another. Your answer should be very short (no more than 3 sentences). You might also find the discussion of the spliceosome and alternative splicing in your textbook helpful.

Any reasonable answer is acceptable. Some examples of acceptable answers include:

- a) The cell can express different proteins that bind to different splice signals on the pre-mRNA to favor one splice configuration over another.
- b) The cell can express an interfering RNA to stop the translation of a certain splice variant.
- c) A splice variant protein may be post-translationally modified (e.g., phosphorylated) to render it active or inactive.

An example of an unreasonable answer is just to say that there is “transcriptional regulation of splice variants.” This statement on its own is not detailed enough and is potentially wrong. A key feature of alternative splicing is that the splice variants are derived from the same pre-mRNA transcriptional product, so they cannot be differentiated at the level of transcriptional initiation.

The ChIP technique has proven to be very valuable, but it can also be quite cumbersome to screen for individual genes in an immunoprecipitated pool of DNA. Fortunately, with the advent of high-throughput sequencers, we can now directly sequence a whole pool of immunoprecipitated DNA. The sequences present in the pool can then be aligned to available genome sequences to determine their origin. This technique is known as ChIP-Seq, and it can be used to determine how a given transcription factor is deployed across the entire genome under given conditions. Such measurements are important for systems-level studies because they provide a more comprehensive view of the multiple nodes in a regulatory network where a particular transcription factor plays a role (e.g., Johnson et al., 2007, “Genome-wide Mapping of In Vivo Protein-DNA Interactions” *Science* 316: 1497-1502).

G. (3 points) For a random DNA molecule, what is the minimum length of sequence you would need in order to map it uniquely to a site in the human genome? Assume that the human genome is 3 billion basepairs in length.

Assuming that the human genome is a random collection of 3 billion bases, a specific sequence of length n can be expected to occur $E = [(3 \times 10^9) \times 4^{-n}]$ times. $E < 1$ when $n = 16$, so you would need a 16 basepair sequence to map it uniquely to a site in the human genome. If students provide an answer greater than 16 basepairs, they should justify what additional assumptions/allowances they made.

H. (4 points) ChIP-Seq experiments normally sequence more than the number of bases you derived in the previous question to guarantee uniqueness of genome mapping. Give one reason

why sequencing the minimum number of bases has a high probability of resulting in redundant genome mappings.

Either of the following answers is acceptable, as well as other reasonable answers:

- a) Variability between immunoprecipitated sequences is much less than random because the DNA probably contains common sequence motifs that the targeted transcription factor binds to. Hence, if you were to only sequence 16 bases, in reality the number of unique bases would only be around 8 bases (because the other 8 or so would represent a common motif), which may not be enough to give a unique mapping to a genome.
- b) The genome itself contains redundancies because many proteins and regulatory elements exhibit homology. These redundancies, which are often flanked by unique sequences, mean that a longer sequence is needed to capture their unique flanks and hence to give a unique mapping.

The figure below shows the results for a ChIP-Seq experiment using an antibody against the human Neuron Restrictive Silencing Factor (NRSF), which is a well-known transcription factor that reduces expression of neuronal proteins in non-neuronal cell types. For this experiment, the ChIP-Seq data are shown as “the frequency of occurrence” (also known as “reads”) of unique neighboring 25-basepair sequences across a particular region of the genome, the NeuroD1 locus (NeuroD1 is a protein that is expressed exclusively in neurons.) This format of the data gives enrichment peaks, as seen in the figure below, based on the location of NRSF binding.

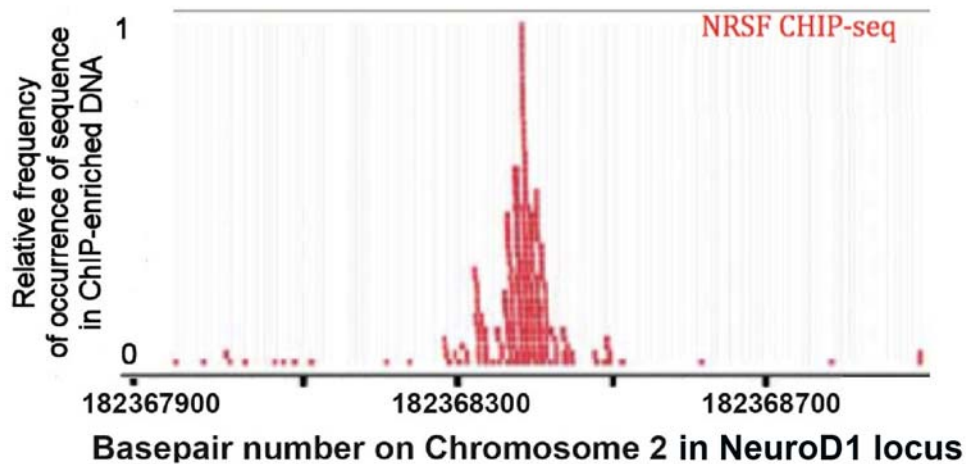


Figure 4: The figure above shows a typical result obtained from performing a ChIP-Seq experiment. In this experiment, the DNA-bound NRSF transcription factor was immunoprecipitated. Co-precipitated DNA was sequenced. A histogram of sequence reads along the NeuroD1 locus in chromosome 2 is shown.

I. (1 point) Using Figure 4, predict the location of the binding site for NRSF on chromosome 2 (*i.e.*, give the approximate basepair number).

The binding site is near basepair number 182368400 (give or take 100 basepairs) because that is where the peak on Figure 4 is.

J. (3 points) The portion of Chromosome 2 shown contains the regulatory region for the NeuroD1 gene. **What conclusion might you draw about the cell type used for this ChIP-Seq assay based on these results and what is known about NRSF and NeuroD1?**

A non-neuronal cell type was used for the ChIP-Seq experiment. NeuroD1 is a neuron-specific gene, and the NRSF transcription factor appears to be reducing or suppressing its expression.

The region of DNA that controls expression of a particular gene (i.e., the **promoter**) is often located within 5-50 basepairs upstream of the gene that is being regulated. However, scientists are increasingly finding DNA control regions (called **enhancers**) that are distant from genes (e.g., up to 10,000 basepairs upstream or downstream of the gene), which are difficult to locate and identify using sequence prediction methods. These enhancer regions play a critical role in the early development of mice, where the differential binding pattern of transcription factors to these enhancers modulates gene expression levels in various tissues. For a given set of known enhancer-binding proteins, a ChIP-Seq experiment can help uncover where these enhancer regions are located and ultimately what genes they might control.

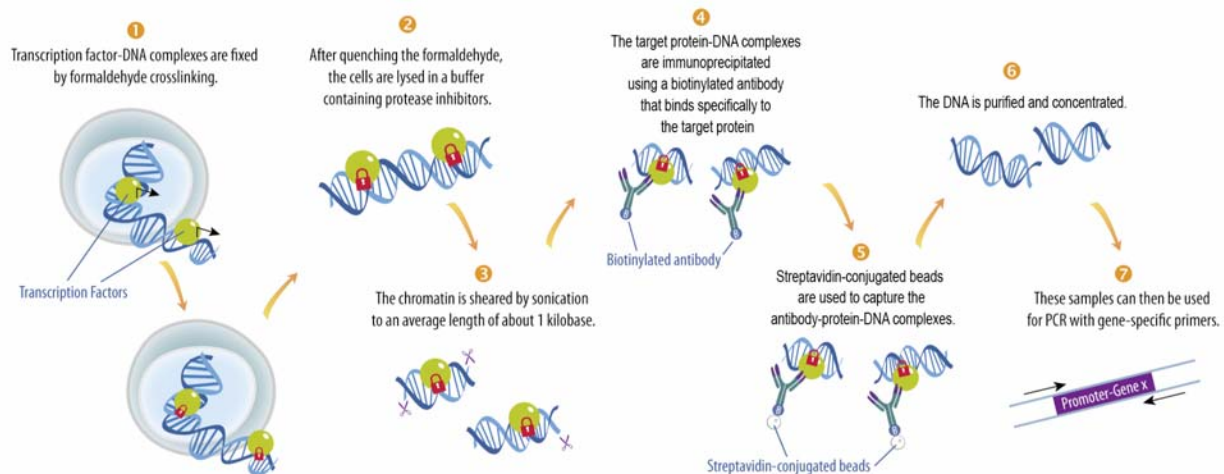
K. (6 points) **How might a transcription factor binding to an enhancer element thousands of basepairs away from the transcription start site still recruit RNA polymerase to the promoter?** Your answer should be short (a few sentences) and can even consist entirely of two well-chosen words.

DNA looping.

Additional material to explain ChIP assays.

A protocol for how a ChIP assay is conducted is diagrammed below. To understand the diagram, you need to know the following:

- i) Formaldehyde is a small molecule that can covalently link macromolecules to each other (i.e., it can crosslink a transcription factor to a piece of DNA).
- ii) To “lyse” a cell means to break open its membranes so as to release the contents of intracellular organelles such as the nucleus.
- iii) Antibodies are proteins produced by the immune system that bind specifically to other proteins (and other macromolecules). They can be raised in an experimental animal to bind almost any protein. Antibodies are usually named according to what they bind, so an antibody against protein X would be called “Anti-X”. Immunoprecipitation is the process by which an antibody that binds to a specific protein is precipitated along with the bound protein. In a ChIP experiment, the result is that a particular transcription factor in a mixture of many other proteins can be precipitated to isolate it and the DNA to which it is bound.
- iv) Biotin is a small molecule that can be covalently attached to a protein so that the biotinylated protein can be captured by streptavidin, a protein that binds very tightly to biotin. When streptavidin is conjugated to beads, the beads can be used to “capture” the biotinylated protein along with anything to which the biotinylated protein is bound.



<http://www.biotechniques.com/biotechniques/protocols/protocolguide/2009/Chromatin-Immunoprecipitation-to-Measure-Transcription-FactorDNA-Interactions/biotechniques-115522.html>

To summarize the diagram, DNA in intact cells is cross-linked to proteins that are attached to it, which would include transcription factors. The cells are then lysed and the DNA is broken into fragments by sonication. Antibodies that bind to the transcription factor of interest are added and DNA-protein-antibody complexes are isolated. DNA is then released from the complex by reversing the cross-linking and the DNA can now be analyzed. This allows researchers to learn whether a particular transcription factor binds close to the promoter or enhancer region of a gene of interest and is therefore likely to regulate it.