7.014 Problem Set 3

Please print out this problem set and record your answers on the printed copy. Answers to this problem set are to be turned in to the box outside 68-120 by 5:00pm on Friday March 16, 2007. Problem sets will not be accepted late. Solutions will be posted online.

1. DNA replication

(a) Why is DNA replication an essential process?

*In order for an organism to grow, its’ cells need to divide. For each round of cell division, DNA has to be replicated such that both the parental cell and daughter cell receive a copy of DNA after division.*

(b) You have created an *in vitro* (in the test tube) DNA replication system using yeast proteins and yeast DNA. One day you accidentally add human DNA polymerase instead of yeast DNA polymerase. You still get DNA replication! Provide an explanation for why human polymerase can substitute for yeast polymerase.

*Both human and yeast polymerase are both eukaryotic polymerases. DNA replication is a highly conserved process. It is possible that the proteins necessary to carry out this process are also highly conserved.*

(c) DNA replication begins at a site along the DNA known as the origin of replication, or ori. Bacterial DNA has one ori, but human DNA has multiple origins. Give one explanation why this is the case.

*Human DNA is so long that if there was only one origin of replication, DNA replication would be too long to be practical. The smallest human chromosome is 10x larger than the E. coli genome. To solve this, humans have evolved to have multiple origins of replication and ensure quicker DNA replication. Additionally, humans have multiple chromosomes (2 sets of 23 chromosomes) whereas bacteria only have one. At the minimum, humans must have one origin per chromosome.*

(d) Below is a depiction of a replication bubble.

(i) Place the following primer on the diagram above: 5’- GCUAUCG -3’ and represent the direction of replication by an arrow.

(ii) Is the primer made out of DNA or RNA?

*RNA*
Question 1 continued

(iii) If the replication fork moves to the right, will the primer be used to create the leading strand of replication or the lagging strand? Explain your answer.

*Lagging Strand.* Since DNA polymerase moves from 5’→3’, and the primer’s 3’ end faces the left replication fork, DNA polymerase can only proceed towards the left. Thus, for the case of the replication fork moving to the right, the direction of replication is opposite of the direction of fork movement, which is consistent with lagging strand replication.

(iv) If you answered lagging strand, explain why this leads to discontinuous replication. If you answered leading strand, explain how this leads to continuous replication.

In lagging strand synthesis, multiple RNA primers need to be made as the DNA at the replication fork is unwound (because the direction of replication is opposite the direction of DNA unwinding). Thus, DNA polymerase must detach from the old primer and re-attach to the new primer during lagging strand replication. Since the polymerase cannot replicate the strand without detaching, lagging strand synthesis is considered “discontinuous”. This is in contrast to leading strand synthesis in which DNA polymerase attaches to one RNA primer and replicates the entire strand without detaching.

2. You have finished your summer internship at the Venter Institute, but you are interested in studying more about the species that you characterized over the summer. Dr. Venter has agreed to let you take the three species that you studied (M, I and T) back to MIT with you so you can investigate them further.

From your initial experiments characterizing how the species obtain energy (Problem Set 1), you noticed that the two autotrophs are capable of surviving in the absence of CO₂ if glucose is provided. This suggests that the autotrophs produce a transporter that can bring glucose from the outside environment into the cell to use the glucose as a source of energy. You scan the genome of Species T (which you sequenced while at the Venter Institute) and find a gene that is homologous to glucose transporters from other species. You call the gene *glcP*.

Further study shows that *glcP* is in an operon with another gene, which you call *glcX*. You are interested in studying the regulation of the *glc* Operon.

You have isolated several mutants (A through E) that are altered for their ability to survive in the presence of glucose and the absence of CO₂. You have an assay for the level of the protein products of *glcP* and *glcX*: GlcP and GlcX, respectively. The results from studying each of the mutants in the assay are presented below. In the following charts, + indicates a wild-type sequence and – indicates a mutant sequence. The results for the haploid strains are shown below.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>+ inducer</th>
<th>- inducer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wild-type</td>
<td>high</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>A</td>
<td>low</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>low</td>
<td>low</td>
</tr>
<tr>
<td>B</td>
<td>low</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>low</td>
</tr>
<tr>
<td>C</td>
<td>high</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>high</td>
</tr>
<tr>
<td>D</td>
<td>high</td>
<td>Low</td>
</tr>
<tr>
<td></td>
<td>low</td>
<td>low</td>
</tr>
<tr>
<td>E</td>
<td>high</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>high</td>
<td>high</td>
</tr>
</tbody>
</table>

(a) What molecule would be an appropriate inducer for the *glc* Operon? 
*Glucose*
(b) Based on your results, you think that the glc Operon is regulated by repression. What observation(s) support this theory?

*Mutants E and C support this theory.*

*If this operon is regulated by repression, there is a repressor that binds to an operator when there is no inducer present. In the presence of inducer, the repressor will no longer bind the operator, and protein levels of GlcX and GlcP will be high.*

Thus, if there are any mutations in the repressor or operator, we will see constitutive levels of GlcP and GlcX both in the presence and absence of inducer, as we see in mutants E and C. Mutations causing constitutive (always “on”) phenotypes are characteristic of operons regulated by repression.

*(In the case of operons regulated by positive regulation, **most** mutations will result in uninducible phenotypes, **not** constitutive phenotypes)*

(c) Based on your knowledge of the lac operon (another operon regulated by repression), you assume that there could be mutations in the promoter of the glc operon, the operator of the glc operon, the glcP or glcX gene sequences, or the gene that produces the repressor of the glc operon. Therefore, which mutants could have mutations in (note: there could be more than one answer):

(i) the coding region of glcP?
   *Mutant B*

(ii) the coding region of glcX?
   *Mutant D*

(iii) the promoter of the glc operon?
   *Mutant A*
   *Mutant C, E (If promoter mutation allows RNA polymerase to bind even in presence of repressor)*

(iv) the coding region of the repressor?
   *Mutant C, E*
   *Mutant A (If the mutation creates a repressor that constitutively binds to operator)*

You then create the following diploids and perform the assay for detecting levels of GlcP and GlcX again.

<table>
<thead>
<tr>
<th>diploid</th>
<th>Genotype</th>
<th>+ inducer</th>
<th>- inducer</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>B⁺D⁻ / B⁺D⁺</td>
<td>High</td>
<td>low</td>
</tr>
<tr>
<td>2</td>
<td>A⁺B⁻⁻ / A⁺B⁻⁻⁺</td>
<td>Low</td>
<td>low</td>
</tr>
<tr>
<td>3</td>
<td>C⁺B⁻⁻ / C⁺B⁻⁻⁺</td>
<td>High</td>
<td>low</td>
</tr>
<tr>
<td>4</td>
<td>E⁺B⁻⁻ / E⁺B⁻⁻⁺</td>
<td>High</td>
<td>high</td>
</tr>
</tbody>
</table>
(d) Explain the results for diploids 1 through 4 based on their genotypes.

Diploid 1:

Diploid 1 shows that the mutations in B and D are recessive and can complement each other (i.e. levels of GlcP and GlcX return to wild-type levels). These mutations are most likely to be in GlcP and GlcX, respectively.

Diploid 2:

In Diploid 2, the wild-type copies of B and D are on same piece of DNA as the mutation from mutant A. The mutations from B and D are on the same piece of DNA that has a wild-type copy of whatever is mutated in mutant A. The mutation in A results in an uninducible phenotype in the haploid organism. The presence of a wild-type A can not complement the mutant A, therefore, the piece of DNA mutated in mutant A acts is a section of DNA that is necessary for regulation (i.e. – not the coding region). The mutation in mutant A is most likely a mutation in the promoter of GlcP and GlcX. A wild-type promoter upstream of GlcX and GlcP can not restore wild-type levels and the phenotype remains uninducible.

Diploid 3:

In Diploid 3, wild-type copies of B and D are on same piece of DNA as the mutation from mutant C, while the mutation from mutant C is on the same piece of DNA as the wild type B and D. Since we see an inducible (wild-type) phenotype for this diploid, this tells us that the mutation in C is most likely in the repressor.

It might be easier to think about what outcome would predict this type of data: A mutation in a repressor or an operator (that prevents repressor binding) in a haploid organism would result in high levels of gene expression. If the mutation is in the operator, you can not restore wild-type regulation with by having a wild-type operator on the same piece of DNA as the mutant genes for GlcP and GlcX. If the mutation is in the repressor, you can restore wild-type levels of regulation if the wild-type repressor gene is on the same piece of DNA as the mutant GlcP and GlcX genes because the wild-type repressor can still be produced and it can bind to the piece of DNA that has wild-type copies of the GlcP and GlcX genes.

Because Diploid 3 shows wild-type regulation in the presence of the mutation from C (only one copy), the mutation must be in the repressor.

Diploid 4:

In Diploid 4, wild-type copies of B and D are on same piece of DNA as the mutation from mutant E, while the mutation from mutant E is on the same piece of DNA as the wild type B and D. Since we see a constitutive phenotype for this diploid, this tells us that the mutation in E is most likely in the operator. Also see the explanation for Diploid 3.

(e) Which mutant has a mutation in the repressor? 
Mutant C

(f) Which mutant has a mutation in the operator? 
Mutant E
Question 2 continued

You have two additional mutants, F and G, that you know have a mutation in the \textit{glcP} coding region. You sequence these two mutants. Below are the sequences you generate from the wild-type Species T and the two other mutants. The sequences for F and G are only of the region where you found a mutation (base pairs 141-210). The underlined regions in the wild-type sequence denote where RNA polymerase associates with the DNA. Transcription begins at the enlarged and bolded C/G pair at 84bp. The mutations in F and G are denoted by being large, bolded and have asterisks above them.

<table>
<thead>
<tr>
<th>wt</th>
<th>5’-TAACCATTGCATATGCCATTGCAATTGGTTTTGCACACATTTAATTAAGAATTCTAGACGAAGTTTTTTACTTG-3’</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1-----------------------------------------------------------------------------+ 70</td>
</tr>
<tr>
<td></td>
<td>3’-ATTGGTAACGATACGTAGCTCAAAACGTTCGTAAAATTTAATTTAAGAATTCATAGACGAAGTTTTTTTACTTG-5’</td>
</tr>
<tr>
<td></td>
<td>5’-GTAGGATTATTATCATGGGCAACCGAACAAATTCTACGTTTACATGAAATCCCTCTCTCTCCTCTCCCT-3’</td>
</tr>
<tr>
<td></td>
<td>71-----------------------------------------------------------------------------+ 140</td>
</tr>
<tr>
<td></td>
<td>3’-CATCCTAAATAGGCAACCGTTTCGCTGTTGTTAAAGTCGAAATGTACTTGGGAGAGAGAAGGAAGGG-5’</td>
</tr>
<tr>
<td></td>
<td>5’-AACTCTACGGCTAACGGATAGTTGGCTCTGCATGATTGGCTGAGGGGTAGCAGCCCTGGGGGTTCTGGTTG-3’</td>
</tr>
<tr>
<td></td>
<td>141---------------------------------------------------------------------------+ 210</td>
</tr>
<tr>
<td></td>
<td>3’-TTAGATCCGATTGCAATTCAAACAGGAGGACTAAAGGCTTCGCCGAGGCCCCCGTCGTTGGGGGCGCCACAG-5’</td>
</tr>
<tr>
<td>*</td>
<td></td>
</tr>
<tr>
<td>F</td>
<td>5’-ATTATCGCGTAACTAGCTTAACTAATGGTTGCTCTGCATGATTGGCTGAGGGGTAGCAGCCCTGGGGGTTCTGGTTG-3’</td>
</tr>
<tr>
<td></td>
<td>141---------------------------------------------------------------------------+ 210</td>
</tr>
<tr>
<td></td>
<td>3’-TTAATTGCGGATTGCAATTCAAACAGGAGGACTAAAGGCTTCGCCGAGGCCCCCGTCGTTGGGGGCGCCACAG-5’</td>
</tr>
<tr>
<td>*</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>5’-AACTCTACGGCTAACGGATAGTTGGCTCTGCATGATTGGCTGAGGGGTAGCAGCCCTGGGGGTTCTGGTTG-3’</td>
</tr>
<tr>
<td></td>
<td>141---------------------------------------------------------------------------+ 210</td>
</tr>
<tr>
<td></td>
<td>3’-TTAGATCCGATTGCAATTCAAACAGGAGGACTAAAGGCTTCGCCGAGGCCCCCGTCGTTGGGGGCGCCACAG-5’</td>
</tr>
</tbody>
</table>

(f) Circle the region(s) of the wild-type sequence where you think mutant A would have a mutation.

Question 2 continued

(g) Your bench-mate, who works on this project with you, has written down the following sequence as the first 10 nucleotides in the transcript:

\[ 5’ \text{CACCGTTTCG} 3’ \]

Explain two reasons why your bench-mate’s sequence is incorrect.

1. RNA contains U, not T
2. Your bench mate has actually written the 3’->5’ sequence of RNA (Shown in Red)
3. your bench mate did not start at the correct nucleotide
4. your bench mate wrote the nucleotides from the template strand, not the nucleotides that are a part of the mRNA.

(h) What are the first six amino acids of the translated sequence?

\[ \text{Met-Asn-Pro-Ser-Ser-Ser} \]
(i) For mutant F:
   (i) What is the effect of the glcP mutation on the primary level of structure of the GlcP protein?

   \textit{The primary level of structure is changed, since the Ser residue is now changed to Phe.}

   (ii) What might be the effect of the glcP mutation on the tertiary level of structure of the GlcP protein?

   \textit{Since Serine is a polar amino acid and Phe is nonpolar, tertiary interactions such as hydrophilic/hydrophobic clusters may be disrupted.}

   (iii) Explain how this mutation in the glcP gene could result in a loss-of-function of the GlcP protein.

   \textit{If the mutation occurs in the active site, changing the Ser to a Phe may result in a nonfunctional glucose transport protein. Additionally, this change might be out of the active site, but could affect the general structure of the protein so that the protein no longer folds properly.}

(j) For mutant G:
   (i) What is the effect of the glcP mutation on the primary level of structure of the GlcP protein?

   \textit{This mutation results in an early stop codon, thus creating a primary sequence that is much shorter.}

   (ii) What might be the effect of the glcP mutation on the tertiary level of structure of the GlcP protein?

   \textit{Having an early stop codon would probably severely disrupt tertiary structure since the protein has lost so many amino acids, of which some were likely responsible for maintaining tertiary structure. Additionally, all of the catalytic regions of the protein might be lost.}

   (iii) Explain how this mutation in the glcP gene could result in a loss-of-function of the GlcP protein.

   \textit{Having an early stop codon could cause multiple problems- the active site could remain untranslated, the protein conformation may be severely affected etc.}

(k) You find that humans also have a glucose transporter that is very similar to the glucose transporter from Species T. You compare the sequences of the genes and the proteins and find the following:

   \begin{tabular}{|c|c|c|}
   \hline
   & Length of Gene & Length of Protein \\
   \hline
   Species T & 1404 base pairs & 468 amino acids \\
   Human & 2000 base pairs & 477 amino acids \\
   \hline
   \end{tabular}

   Explain why the two genes are different lengths but encode for proteins of almost exactly the same length.

   \textit{Because humans have introns in their genes that are spliced out.}
(l) All of the \textit{glcP} mutants (A-G) that you identified prohibit Species T from being able to survive on glucose in the absence of CO$_2$. Do you think that mutations always have a negative impact on an organism’s life?

\textit{No. One example is evolution, which is actually based on “good” mutations that give an organism a higher chance to survive and reproduce.}

\begin{center}
\textbf{The Genetic Code}
\end{center}

\begin{tabular}{|c|c|c|c|c|}
\hline
 & U & C & A & G \\
\hline
\hline
U & Phe & Ser & Tyr & Cys & U \\
 & Phe & Ser & Tyr & Cys & C \\
 & Leu & Ser & STOP & STOP & A \\
 & Leu & Ser & STOP & Trp & G \\
\hline
C & Leu & Pro & His & Arg & U \\
 & Leu & Pro & His & Arg & C \\
 & Leu & Pro & Gln & Arg & A \\
 & Leu & Pro & Gln & Arg & G \\
\hline
A & Ile & Thr & Asn & Ser & U \\
 & Ile & Thr & Asn & Ser & C \\
 & Ile & Thr & Lys & Arg & A \\
 & Met & Thr & Lys & Arg & G \\
\hline
G & Val & Ala & Asp & Gly & U \\
 & Val & Ala & Asp & Gly & C \\
 & Val & Ala & Glu & Gly & A \\
 & Val & Ala & Glu & Gly & G \\
\hline
\end{tabular}
Peptide bonds
C
O
N
H
C
C
O
H
R1
N
H
C
C
O
N
H
H
H
R2
R3
Protein ... +
C
C
O
H
NH3
C
H
CH3
CH3
-
+
C
C
O
H
NH3
CH3
-
+
ASPARAGINE
(asN)
ARGININE
(arg)
ALANINE
(ala)
GLUTAMIC ACID
(glu)
GLUTAMINE
(glN)
GLYCYCINE
(gly)
CYSTEINE
(cys)
HISTIDINE
(his)
ISOLEUCINE
(ile)
LEUCINE
(leu)
LYSINE
(lys)
METHIONINE
(met)
PHENYLALANINE
(phe)
PROLINE
(pro)
SERINE
(ser)
THREONINE
(thr)
TRYPTOPHAN
(trp)
TYROSINE
(tyr)
VALINE
(val)