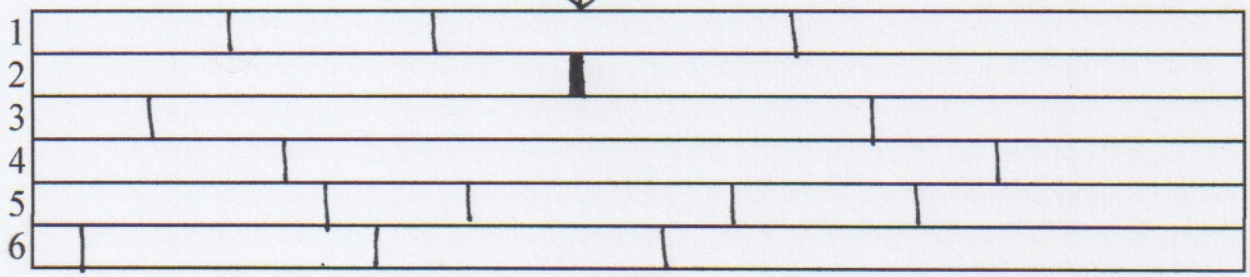
7.03 PSet 3 Solutions

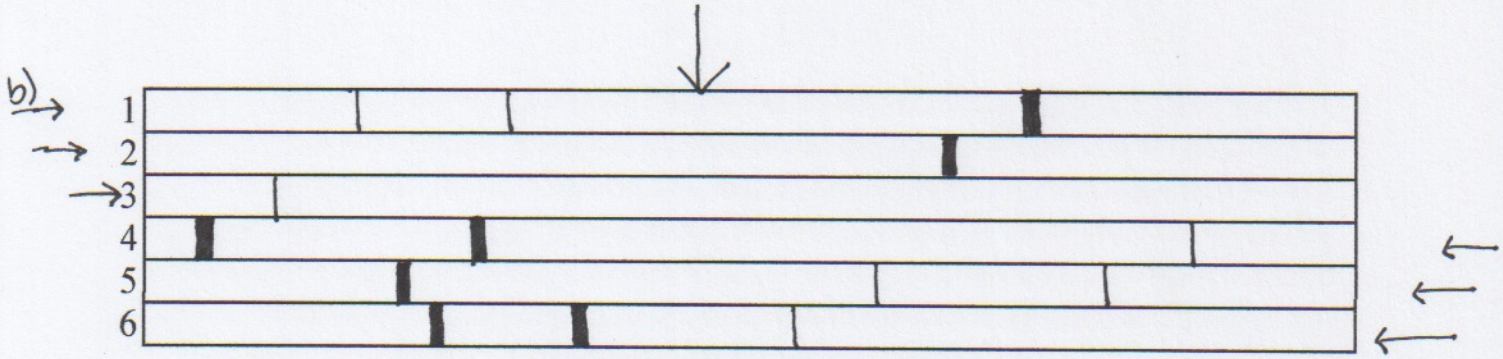
Spring 2011

#1 a) Changes will be in "bold"



This is an amber mutation inserted INTO the GENE ONLY.

b)



The insertion of a single bp doesn't destroy stop codons, only moves them from 1 reading frame to another.

For example, ORIGINAL sequence ~ TAG stop codon in frame 1:

Frame 1: att att att TAG att

Frame 2: a tta tta ttT AGa tt

Frame 3: at tat tat tTA Gat t

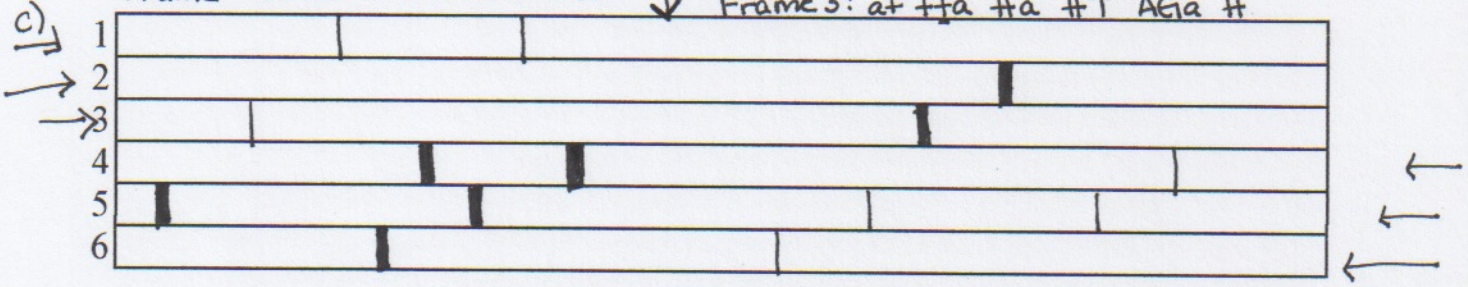
Then, insert 1 "t" base after 1st 3 bases

Frame 1: att tat tat tTA Gat t

Frame 2: a ttt att att TAG att

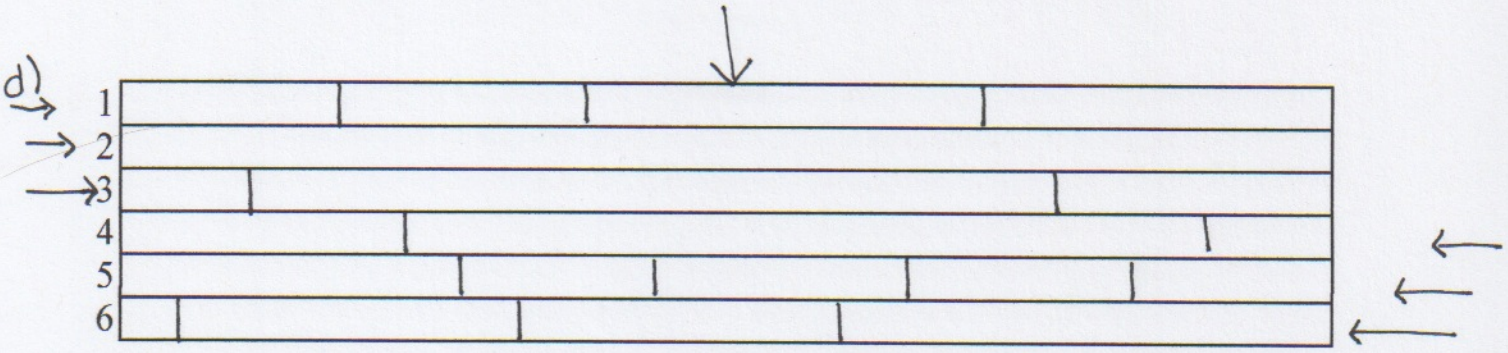
Frame 3: at tta tta ttT AGa tt

c)



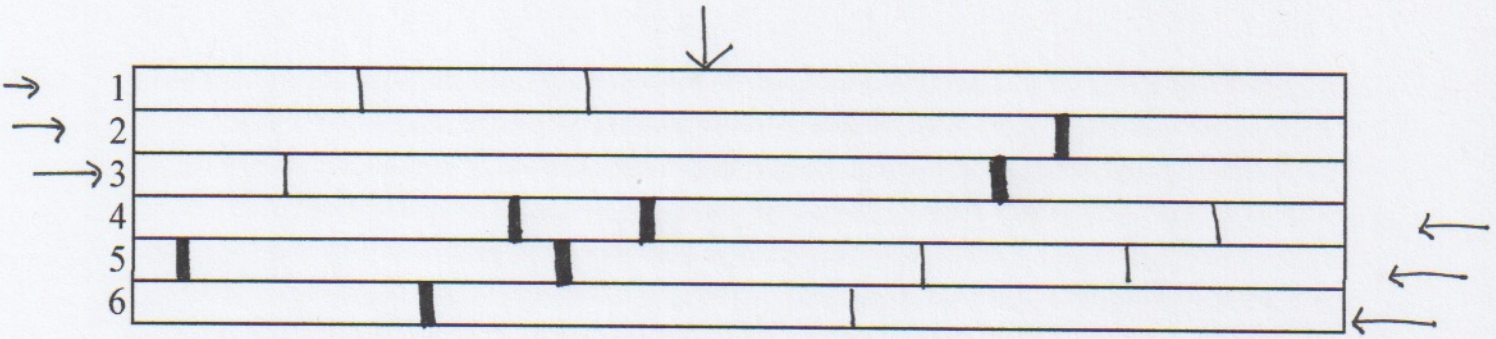
Again, a 2 bp insertion doesn't destroy stop codons, only alters the frame. You can replicate the same example as above for proof if needed.

1 d)



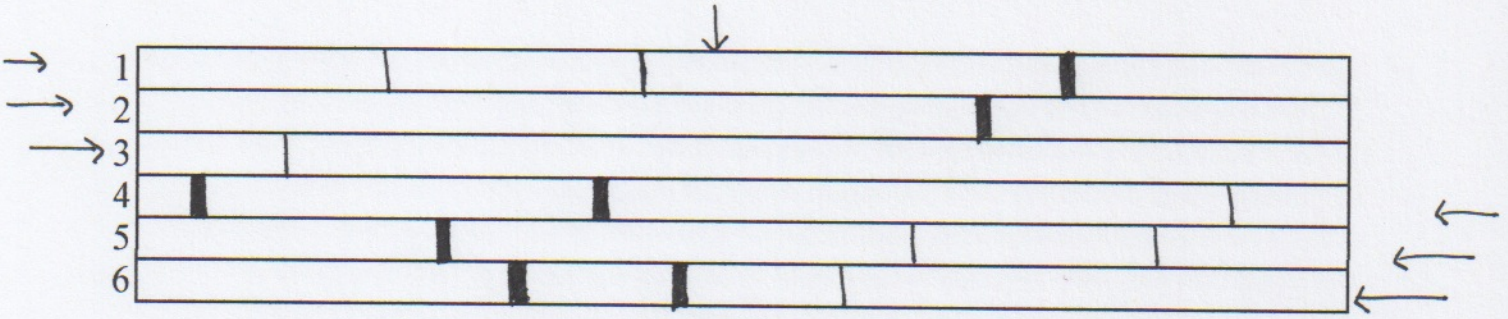
An insertion of 3 base pairs adds a full codon, ∴ stop codons remain in the original reading frame.

1e)



Deletion of a single bp. Shifts stop reading frame.

1f)



Deletion of 2 bp's. Shifts stop reading frame.

FINAL ANSWERS SHOULD BE A FUNCTION OF A·T CONTENT!!! PSET 3 Soln's.

2 a) In 50% A·T sequences, the probability of getting any 1 stop codon is the prob. of getting any 3 bases in a row.

$$P(\text{any base})^3 = \left(\tfrac{1}{4}\right)^3 = \tfrac{1}{64}$$

Since there are 3 stop codons total (UAA, UAG, UGA)

$$P(\text{any stop}) = 3\left(\tfrac{1}{64}\right) = \tfrac{3}{64}.$$

$\therefore$ For UAA: $p(UAA) = \left(p\,\frac{(A\cdot T)}{2}\right)^3$

For UAG or UGA: $p(UAG) = \left(p\left(\frac{A\cdot T}{2}\right)\right)^2\left(p\left(\frac{GC}{2}\right)\right) = \left(p\left(\frac{A\cdot T}{2}\right)\right)^2\left(\frac{1-p(AT)}{2}\right)$

$$\Rightarrow \boxed{P(\text{stop}) = \left(p\left(\tfrac{AT}{2}\right)\right)^3 + 2\left(p\left(\tfrac{AT}{2}\right)\right)^2\left(\frac{1-p(AT)}{2}\right)}$$

For example, if an organism had an AT content of 30%,

$$P(\text{stop}) = \left(\tfrac{.30}{2}\right)^3 + 2\left(\tfrac{.30}{2}\right)^2\left(1-\tfrac{.30}{2}\right) = (.15)^3 + 2(.15)^2\left(\tfrac{1-.30}{2}\right)$$

$$= 0.019, \text{ or } 1.9\%.$$

b) P(ORF of @ least 200 codons) = p(NOT getting a stop for @ least 200 codons).

From a) $p(\text{STOP}) = \left(p\left(\tfrac{AT}{2}\right)\right)^3 + 2\left(p\left(\tfrac{AT}{2}\right)\right)^2\left(\frac{1-p(AT)}{2}\right)$

$$p(\text{not stop}) = 1 - p(\text{stop})$$

$$p(\text{not stop for } n \text{ codons}) = (1-p(\text{stop}))^n$$

$$\boxed{\left[1-\left(\left(p\left(\tfrac{AT}{2}\right)\right)^3 + 2\left(p\left(\tfrac{AT}{2}\right)\right)^2\left(\frac{1-p(AT)}{2}\right)\right)\right]^{200}}$$

@ A·T = 30%, For 200 codons, $p(200 \text{ codon orf}) = [1-0.019]^{200} = [0.980875]^{200} = \boxed{0.021, \text{ or } 2.1\%}$

c) First, calculate the total number of random ORFs predicted for a 5 Mbp genome. To do this, get the total # of codons N, which is the length of the genome times the 6 reading frames, divided by 3 bases per codon. Then, determine the total # of fortuitous ORFs. To do this, take N and divide by the average length of each ORF given the A·T content. Since the probability from part a is the inverse of the avg. length of each fortuitous ORF, we can multiply by this function. Finally, multiply by the prob. of finding a given length ORF from pt. b. →

$\Rightarrow$ Final equation is: $\boxed{N(p)(1-p)^n}$, the prob. of a random ORF in the genome.

where N = total # of codons $= \frac{(6 \text{ reading frames})(5 \text{ Mbp})}{3 \text{ bases per codon}} = 10{,}000{,}000$

p = probability of a stop codon $= \left(p\left(\frac{AT}{2}\right)\right)^3 + 2\left(p\left(\frac{AT}{2}\right)\right)^2\left(\frac{1-p(AT)}{2}\right)$

and n = length of a fortuitous ORF.

So, for our 30% AT example, for fortuitous ORFS of ≥ 200 codons,

$$\text{prob} = 10{,}000{,}000\,(p(\text{stop}))(1-p(\text{stop}))^{200}$$

Substitute from parts a & b:

$$\text{prob} = 10{,}000{,}000(0.019125)(0.980875)^{200} = 4{,}021 \text{ fortuitous ORFs in a 30\% AT genome.}$$

$\Rightarrow$ porportion of false positives $= \frac{\#\text{ fortuitous orfs}}{\text{total orfs}}$

$$\text{prop false positives} = \boxed{\frac{N(p)(1-p)^n}{N(p)(1-p)^n + (\#\text{actual genes})}}$$

Contuing with our 30% AT example, prop. false positives $= \frac{4{,}021}{4{,}021+4{,}000}$

$$= \frac{4{,}021}{8{,}021} \approx 50.1\% \text{ false positives}$$

$\Rightarrow$ You would rather search in a relatively high AT content organism, as there would be fewer fortuitous ORFs.

3. Tryptophan 5'-UGG-3' (RNA)

⇒ anti codon loop is 5'-CCA-3'

so encoding DNA must be

5'-CCA-3'
3'-GGT-5' ← template strand

Amber mutation 5'-UAG-3' (RNA)

anti codon would be 5'-CUA-3'

encoding DNA

5'-CTA-3'
3'-GAT-5' ← template strand

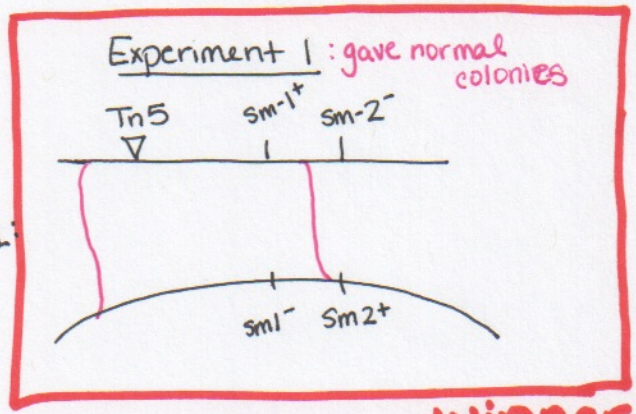
4. a) $\text{Dist}_{Tn5\text{-}sm1}: \frac{10 \pm \sqrt{10}}{50} = 0.20 \pm 0.063 = \boxed{20 \pm 6.3\%}$

b) $\text{Dist}_{Tn5\text{-}sm2}: \frac{9 \pm \sqrt{9}}{50} = 0.18 \pm 0.06 = \boxed{18 \pm 6\%}$

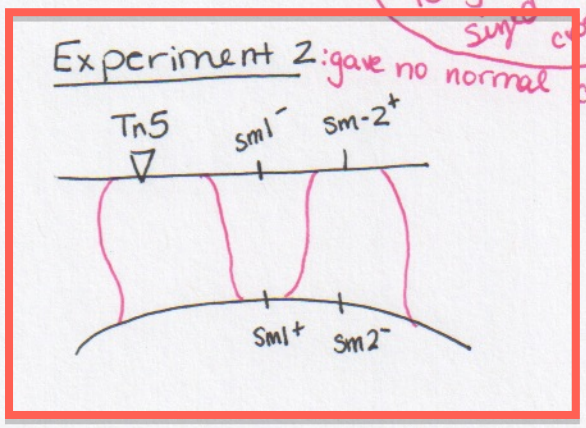
No, from this data you cannot order the two genes and the Tn5 insertion, as the error is too large, and we don't even know if sm1 and sm2 are on the same side of the Tn5 insertion.

c) Since normal sized colonies form more frequently in the first experiment, we can deduce that they are the outcome of a double cross-over event. The second experiment produces no normally-sized colonies ∴ would most likely require a quadruple crossover event.

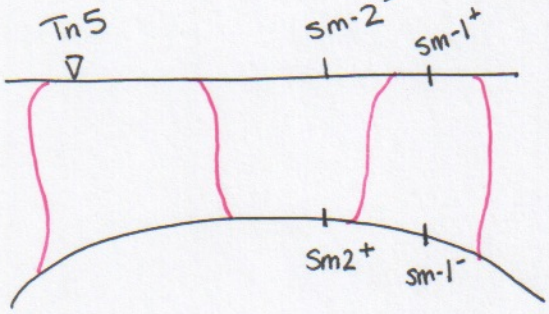
To get a normally-sized colony: 2 crossovers shown



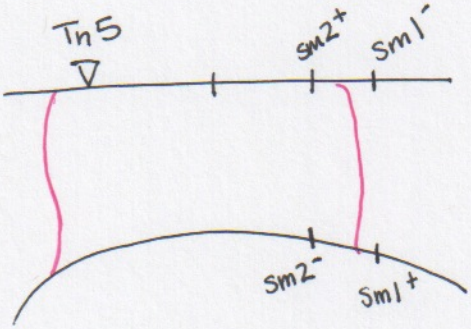
ORDER 1:



winner !!



ORDER 2:



⟹ MAP: Tn5 —— sm1 —— sm2
Tn5 ⟷ sm1: 20±6.3%
Tn5 ⟷ sm2: 18±6%

Nok, no distance between sm1 & sm2 can be shown! Cotransduction curve isn't linear.

Further explanation of 4c)

As usual for a three-factor cross, always look for the rarest class. In this case, this results from experiment 2, which gave NO normal-sized colonies. Next, looking at the two possible orders, distinguish the one which requires the most rare recombination event (in this case, the quadruple crossover of order #1).

4 d) The first experiment gives:

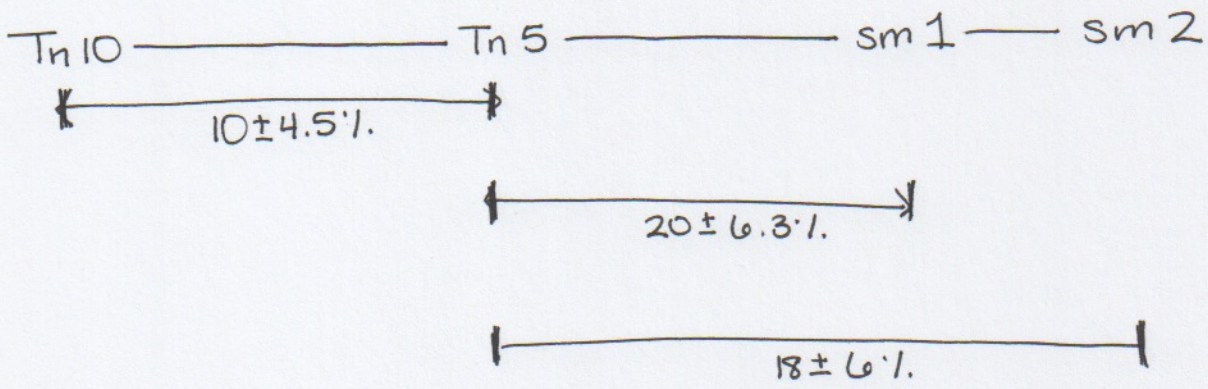
$\text{Dist}_{\text{Tn10}\cdot\text{sm1}} \approx 0\%$ cotransduction (unlinked to sm 1)

The second experiment gives:

$$\text{Dist}_{\text{Tn5-Tn10}} = \frac{5 \pm \sqrt{5}}{50} = 0.10 \pm 0.045 = 10 \pm 4.5\%$$

$\Rightarrow$ most likely map is:

note: not to scale.



5. Notice, that when integrated, the origin of transfer points in the opposite direction of the arrows for the IS sequence. Therefore, by convention it will first transfer loci in the direction the IS sequence arrow is pointing.

To isolate the appropriate Hfr strain, I would perform interrupted mating experiments on several different Hfr strains mated to my mutant strain. I would want to wait a short period of time, then halt the mating and select for C+ mutant strains. I would then screen these colonies by replica plating to make sure that they were B- and A-. This would ensure that the F plasmid inserted into the third IS sequence that is closest to C. The secondary screening is necessary because insertion into the first IS sequence would pass both B+ and C+ relatively early in the mating.

Once the correct Hfr strain has been isolated, I would grow up the strain to allow the plasmid to recombine out of the genome. Then, I would perform another mating experiment. The experiment would again be stopped early. This time I would only need to select for B+ mutant strains to find the F' strains that had looped out of the genome taking B with them. Strains that had not recombined out with B+ would pass B+ only after a very long period of mating or not at all.

You should also replica plate onto A- and C- media and screen for death. This ensures the F' has ONLY B (not A or C).

Please keep in mind that this represents one possible solution (where the sequence recombined with the third IS sequence closest to C). This could also be done with two other IS sequences on the chromosome.