7.03 Exam 3 Review Problems - SOLUTIONS

Problem 1

There is a rare recessive X-linked trait that affects 1/8000 males in a rabbit population. The mutation causes black spots on the white rabbits (normal phenotype is plain white).

(a) What is the value of q?

Since males are haploid at this gene locus, f(affected males) = q = 1/2000 = **0.0005**

(b) In what proportion of matings would this trait affect half of the male and female offspring?

To affect half of the offspring, mother must be a carrier and father must be affected.
p = 1 – q = 1 – 1/2000 = 0.9995
P(mother is carrier) = 2pq = 2*(0.9995)*(0.0005) = 0.001
P(father is affected) = q = 0.0005

P(half of children affected w/ trait) = P(mother is carrier)*P(father is affected) = 0.001*0.0005 = **$5x10^{-7}$**

(c) Now assume only 20% of the affected rabbits survive. What is the value of S?

Since 20% survive, the fitness is 20% or 0.20.
S = 1 – fitness = 1 – 0.20 = **0.80**

(d) By how much would the allele frequency change between the current generation of rabbits and the next generation?

Currrent q = 0.0005
$\Delta q = -Sq/3 = -(0.80*(0.0005)/3 =$ **-0.000133**

Thus, it would decrease by 0.000133.

(e) Now new mutations are introduced to the population, with a mutation rate of $5x10^{-3}$. What would the allele frequency q equal after a new steady state had been reached?

$\mu = 5x10^{-3}$, S = 0.80
At steady state, $\Delta qsel + \Delta qmut = 0$, so $q = 3\mu/S = 3*(5x10^{-3})/0.80 =$ **0.01875**

Problem 2

For a specific gene, there are only two possible alleles, A and a. Professor Regev makes a new population where f(A) = f(a) = 0.5. Assume no new mutations, heterozygous advantage or selective disadvantage.

a) According to the conditions stated in the problem will this population most likely be in Hardy-Weinberg Equilibrium or in a balanced polymorphism? What are the frequencies of the possible genotypes?

Since there are no new mutations, heterozygous advantage or selective disadvantage, this means that the population is most likely in H-W Equilibrium.

$f(AA) = (0.5)^2 = 0.25$
$f(Aa) = 2*(0.5)^2 = 0.5$
$f(aa) = (0.5)^2 = 0.25$

b) Let's say an individual with the genotype AA or Aa is normal, but someone who is aa unfortunately dies shortly after birth. What is s, the selective disadvantage? After an infinite number of generations, what will be the allele frequencies for A and a? Explain.

s =1
After an infinite number of generations, f(A) = 1 and f(a) = 0. This is because there is a negative force that acts against allele a. Homozygous individuals aa die shortly after birth and these alleles are correspondingly eliminated from the population. In addition, you can say that the selective disadvantage is a negative force acting against allele 'a,' and that there is not positive force that is helping it "come back" into the population.

c) There has been a new vaccine that lets 40% of the individuals who are aa survive. What is the new s? Will f(A) and f(a) be the same as in part b after many generations? Explain. If they do change, what are the new frequencies?

s = 0.6
The allele frequencies will be the same as in part b. This is because there is still anegative force acting against allele 'a,' and no positive force helping it come back into the population. However, the equilibrium frequencies will be reached at a slower rate since 60% of the individuals who are aa survive, but in part b, all of them die.

d) Now assume that there can be mutations, where 'A' is mutated to 'a' at a rate of $\mu = 10^{-5}$. Assuming that we start from the beginning population where f(A) = f(a) = 0.5 and the vaccine is there, what will be the equilibrium allele frequencies of p and q?

In this example, we have a selective disadvantage (0.6) and a mutation rate (10^{-5}). The selective disadvantage eliminates allele 'a' from the population, but the mutation rate adds the allele to the population. This creates a balanced polymorphism. Since the disease is autosomal recessive, the equilibrium frequencies are:

q = sqrt(μ/s) = 0.0041

p = 1 – q = 0.996

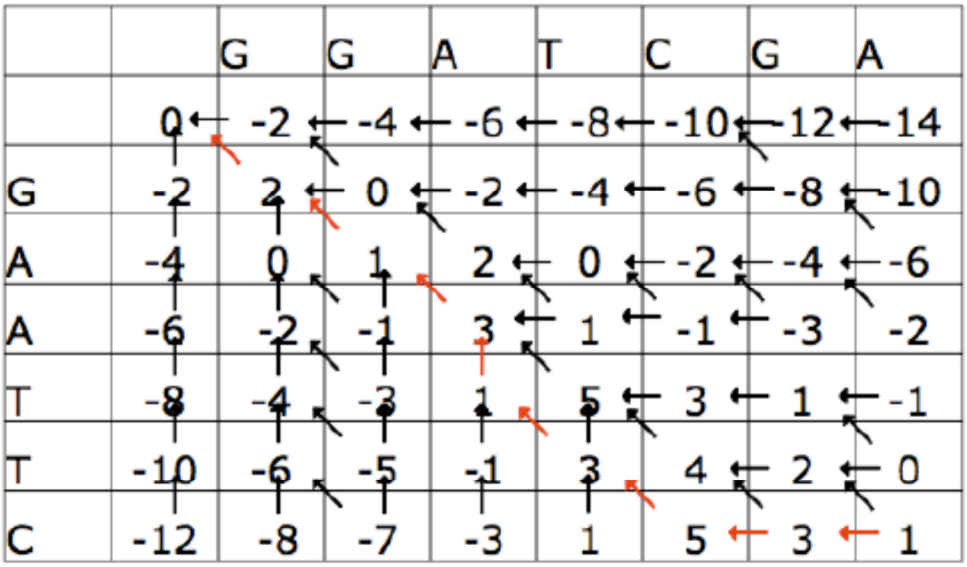
Problem 3

You have two sequences that you want align
A. Using the Needleman-Wunsch algorithm to perform a global alignment of the following two nucleotide sequences:

GAATTC
GGATCGA

And using the following scoring scheme:
Match = +2
Mismatch = -1
Gap = -2 per position

You get the following Dynamic Programming (DP) matrix with an optimal path marked in red.



What is the alignment indicated by this matrix?

```
GAATTC--
|  |  ||
GGA-TCGA
```

What is the score for this alignment?

Number of matches = 4
Number of mismatches = 1
Number of gaps =3

Score=(4*2)+(1*-1)+(3*-2)=1

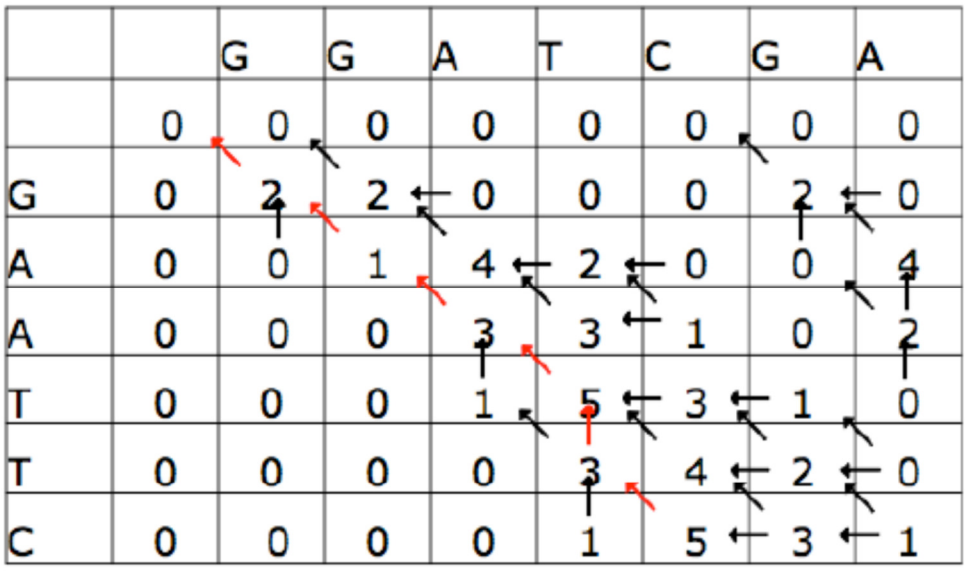
What is the score of the following alignment using the above scoring rules?

```
GAATTC---
GGAT--CGA
```

Number of matches=3
Number of mismatches=1
Number of gaps=5

Score=(3*2)+(1*-1)+(5*-2)=-5

B. Using the Smith-Waterman algorithm to perform a local alignment of the above nucleotide sequences using the same scoring scheme you get a DP matrix with an optimal path marked in red that looks like this



		G	G	A	T	C	G	A
	0	0	0	0	0	0	0	0
G	0	2	2	0	0	0	2	0
A	0	0	1	4	2	0	0	4
A	0	0	0	3	3	1	0	2
T	0	0	0	1	5	3	1	0
T	0	0	0	0	3	4	2	0
C	0	0	0	0	1	5	3	1

What is the alignment indicated by this matrix?

```
GAATTC
|  ||  |
GGAT-C
```

What is the alignment score?

Matches=4
Mismatches=1
Gaps=1

Score=8-1-2=5

C. What is the difference between the local and global alignments in this example?

The global alignment uses all bases and the local alignment only aligns some of the bases. The effect is that the local alignment gives a better but shorter alignment of the two sequences.

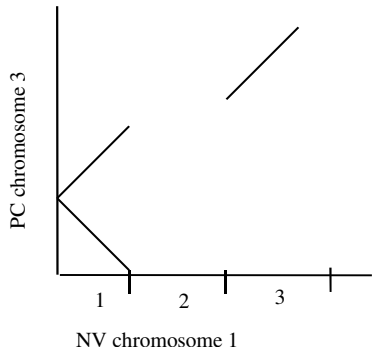
D. You find two proteins that you suspect have a common protein domain because they both can bind to the same DNA sequence. You wish to determine what this protein domain is that allows them to do this. Having honed your alignment skills as a 7.03 student you decide to align the two sequences and look for regions of high similarity. What kind of alignment (global or local) would you perform? Explain.

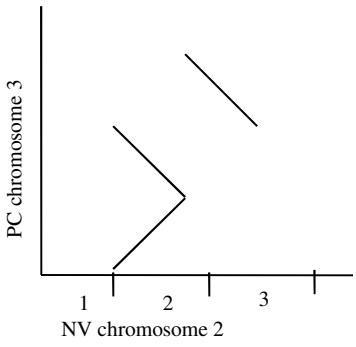
You'll want to perform a local alignment to find a region within the the protein that matches well. The reason you want a local alignment is because you don't care (nor expect) most of the protein to match between the two but you are interested in a specific region that you expect should be highly similar.

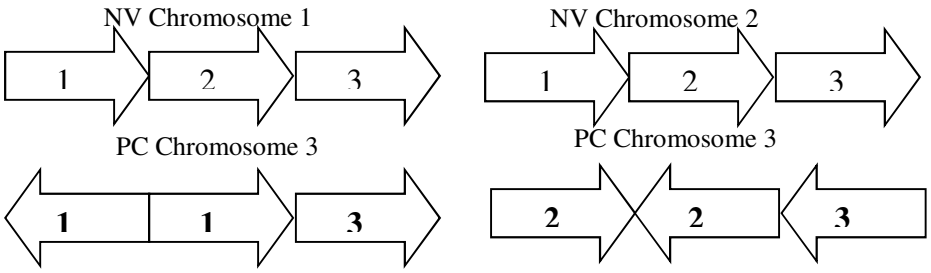
Problem 5

You are attempting to identify coding regions in a newly sequence genome of a sea anemone species *Protantheae carlgren* (PC). At your disposal is relatively well annotated genome of a model organism, the starlet sea anemone, *Nematostella vectensis* (NV).

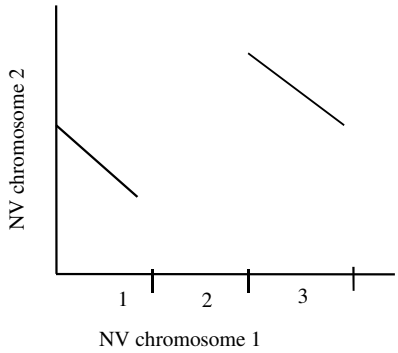
a. You begin with a large-scale chromosomal comparison between the two species. Underneath each dot plot, draw in the PC chromosome 3 arrangement relative each NV chromosome given. Use arrows to indicate sequence direction and number the segments accordingly (1, 2, 3).







b. Based on the information in the graphs in part a, fill in the following dot plot comparing NV chromosomes 1 and 2.



c. You discover three sequence segments in PC with similarity to **start** of a known protein coding region in NV (indicated by red text). Based on the alignments shown, choose the PC sequence that most likely represents truly homologous, viable coding region and state your reasoning. Note that the non-template DNA strand is given for each sequence so consider only forward reading frames, and it is **not** necessary to translate any sequence.

Unaligned NVsequence:

```
5′ CCATGACCTTCGACTCAGTCATCACTCTTGATGAT
     (start)
```

PC sequence 1 alignment:

```
NV: 5′CCATGACCTTC__GACTCAGTCATCACTCTTGATGAT
PC: 5′CCATGACCTTCGGGACTGA_TCATCACTCTGCTTGAT
      ***********  **** * **********   ****
```

PC sequence 2 alignment:

```
NV:  5′CCATGACCTTCGACTCAGTCATCACTCTTGATGAT
PC:  5′C_ATGACGTT_TAC__AGTAATAACCCTAGACGAT
       * ***** **  **  ****** ** ** ** ***
```

PC sequence 3 alignment:

```
NV:   5'CCATGACC_TTCGACTCAGTCATCACTCTTGATGAT
PC:   5'CCATGAGCTTTC_ACGCAGACTTGTCTCCTCATGTT
        ******** *** ** *** ***  ***** *****
```

Sequence 2 contains the most likely conserved coding region. Sequence 1 is highly similar at the nucleotide level, but it can easily be eliminated because it contains an early frameshift mutation that is never corrected. Sequences 2 and 3 both maintain the proper reading frame through the majority of the given sequence. They also have roughly the same number of SNPs. However, sequence 3 has a high proportion of SNPs in the first and second codon positions, while sequence 2 has SNPs almost exclusively in the third 'wobble' coding position. Therefore, sequence 3 encodes a primary amino acid sequence that more similar to the NV protein coding region.