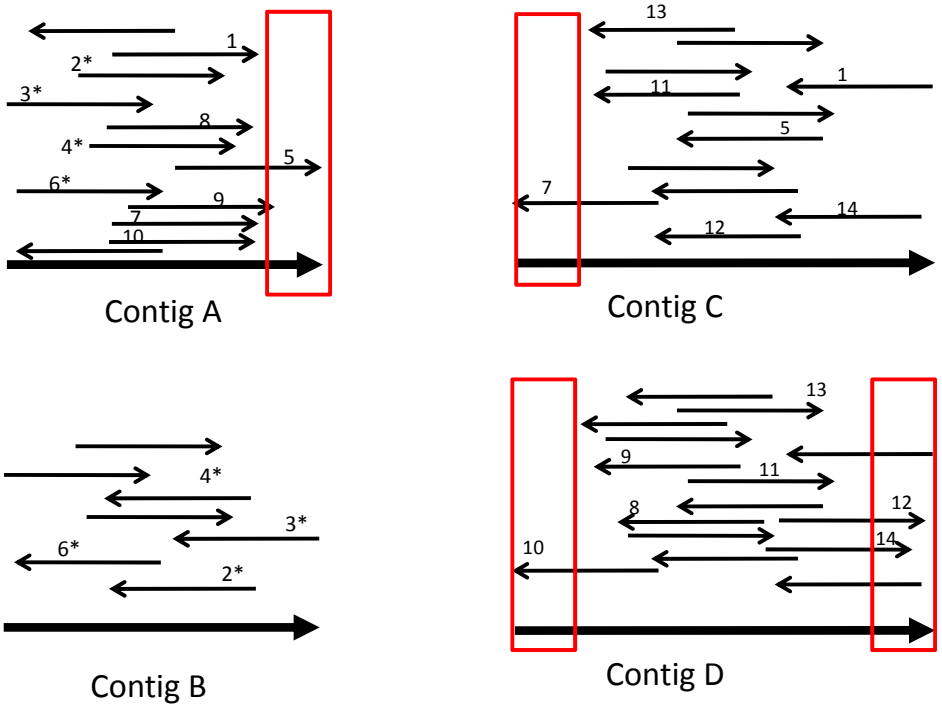
7.03 Problem Set 5

Due before 5 PM on Friday, April 22

Hand in answers in recitation section or in the box outside of 68-120

1. You are assembling the genome of a new species, relying on shotgun sequencing. The sequencing was done with two types of clones. One class was made with inserts approximately 2-4Kb in size, and another with much longer inserts, about 100Kb in size. By identifying regions of overlap, you have assembled four contigs, A, B, C, D from the reads you obtained.



Each contigs is ~1Kb long. In the plots below, the reads (and their orientations) are shown as thin black arrows, and reads from the same clone have the same number (reads with no numbers do not have a mate pair). Reads from clones with 100Kb (long) inserts are marked with an asterisk (*). A short interspersed repeat sequence present in millions of copies in the genome is marked by a red box whenever it appears in a contigs.

a) Examine contigs A. Do you agree with the inclusion of the repeat sequence within the contigs? Why or why not?

Yes, because we have overlapping reads with sequences outside of the repeat. Such as with 5 and 8 in contig A.

b) Draw the relative organization of contigs A and B. Approximately, how far apart do you expect their sequences to be in the genome? Explain your answer.
Because we have paired end reads between A and B only in the 100kb (*)constructs, A and B must be about 100kb apart. The paired end reads in A are facing to the right, while the paired end reads in B are facing to the left, indicating B is to the right of A.

A > ←-------~100kb-----→ B >

c) Now consider contigs C. Draw its position relative to contigs A and B, explain your answer. Approximately, how far apart is contig C from contig A? from contig B?
Because C and A have paired end reads (1,5, and 7) we can assume they are 2-4kb apart. B shares no paired end reads with C, and therefore is likely under 100kb from B.

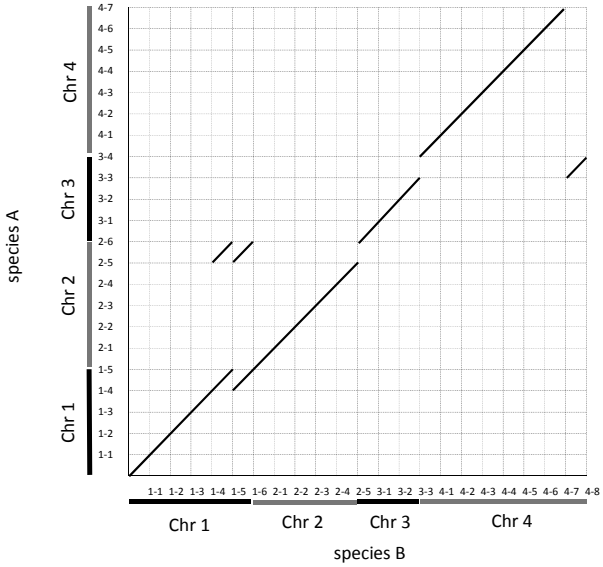
A > ←-------~2-4kb-----→ C > ←-------~96kb-----→ B >

d) Finally, consider contig D and draw its position relative to A, B, and C. If you knew that read 5 in contig A and read 10 in contig D perfectly overlap (100% sequence identity) only within the region marked with the red box, would it affect your answer? If so, how? If not, why not?
D has paired end reads with both C and A. The direction of the paired end reads tells you that D is in between A and C. So A, D, and C must all be within 4kb of each other. ~2kb would likely be less than 2kb because we also need to consider the size of the contigs themselves.

A > ←-------~2kb-----→ D > ←-------~2kb-----→ C > ←-------~96kb-----→ B >

If the only overlap between read 10 in D and read 5 in A was in the red box (repeat), it would not affect our answer. It is very possible that the two contigs do not even share the same repeat region, and rather the red box in D and A represent two different instances of that repeat in the genome. This is the way it has been drawn in the above diagram.

e) After you finished assembling the genome of species A to your satisfaction, you compare it to that of species B, using a dot plot (below, chromosome numbers in each species are marked by alternating black and grey bars, with appropriate coordinates). From the analysis of Species A's genome, you found that a gene important for pathogen evasion is encoded in chromosome 2 between coordinates 5 and 6 (2-5 to 2-6).



Is this gene present in species B? If so, in how many copies and in which coordinates?

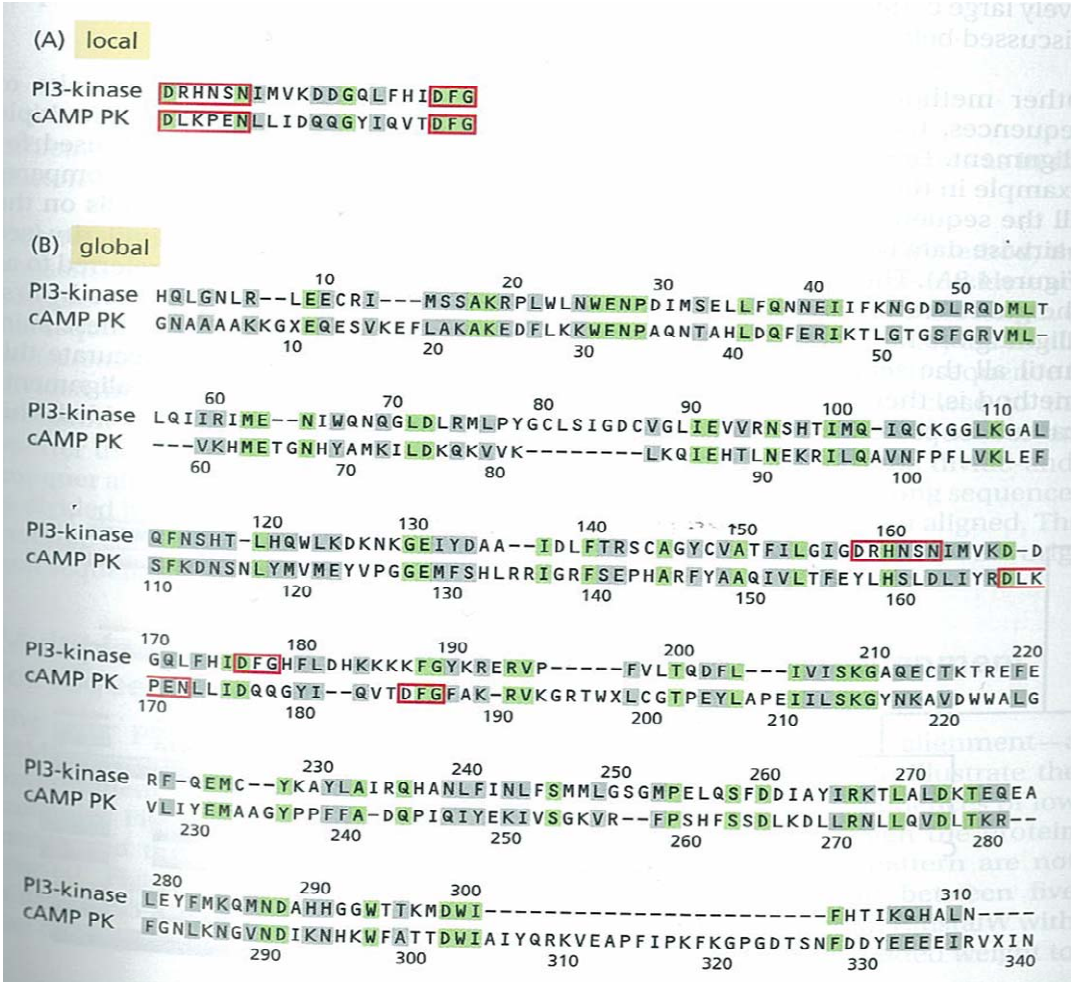
Yes, it is present in two copies in tandem from 1-4 to 1-5 and 1-5 to 1-6.

f) In how many copies is the gene present in species A?

It is present in two copies in species A also.

g) What happened in species B to the sequence at coordinates 3-3 to 3-4 of species A? ***The gene was transposed from 3-3/ 3-4 to 4-7/4-8.***

2. You are studying a human and mouse genes that are somewhat similar. From previous studies, you know that these proteins are often modified by the same enzyme, and are interested in finding the region that may be modified in both by the same enzyme. Such regions tend to be very short. You decide to try both a local and a global alignment, and obtain the following results. Green shading indicates identical amino acids, grey indicates amino acids with chemical properties (such as isoleucine and leucine) and the red boxes mark the regions flanked by conserved residues in the local alignment.



a) Which of the results is more appropriate to pursue for your purposes? Explain?

The local alignment. Because we only want to find a single conserved motif in the protein, we want to use a local alignment to find the single best stretch of conservation between the two proteins.

b) Do the local and global alignments agree on the conserved residues? What could be the reason for a discrepancy?

No they do not. Because the global alignment must find the best possible alignment score using all residues in both proteins, it will invariably match different conserved residues than that of a local alignment.

c) Perform a **local** alignment of the following two peptides from the sequences above

`NIWQGL` and `GNIHQY`

Show your dynamic programming matrix and traceback arrows. Highlight the path of the alignment and write down the optimal **local** alignment. Use the following scores

Match: +2
Mismatch: -2
Gap: -8

		G	N	I	H	Q	Y
	0	0	0	0	0	0	0
N	0	0	2	0	0	0	0
I	0	0	0	4	0	0	0
W	0	0	0	0	2	0	0
Q	0	0	0	0	0	4	0
G	0	2	0	0	0	0	2
L	0	0	0	0	0	0	0

Two possible solutions:

`-`**`NI`**`WQGL`

`G`**`NI`**`HQY`

or

`-`**`NIWQ`**`GL`

`G`**`NIHQ`**`Y`

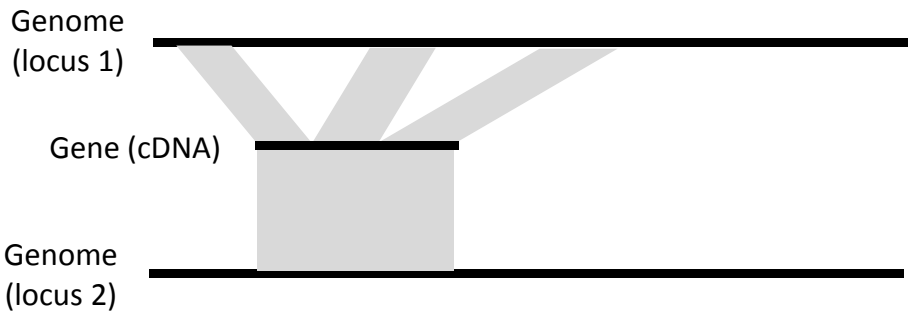
d) How could you revise the mismatch score to reflect what you know about the chemical properties of amino acids?

Since some amino acids have very similar properties (leucine/isoleucine, serine/threonine, etc.), we could consider <u>decreasing</u> the mismatch penalty when an alignment aligns chemically similar residues.

e) Studies from other species indicate that your gene of interest is often present in multiple (similar) copies in other species' genomes. You are interested to explore if your gene of interest was also duplicated in the human genome. Would you use a global or local alignment? Explain.

Global. A global alignment would force the program to consider stretches of DNA that are similar to your entire gene of interest. A local alignment may find very good alignment to a small domain/section of your gene, but would not take into account the rest of the gene's sequence.

f) By comparing your gene's sequence to the rest of the genome you find significantly similar alignments to two loci (besides the gene locus), which appear below. The grey shading indicates the



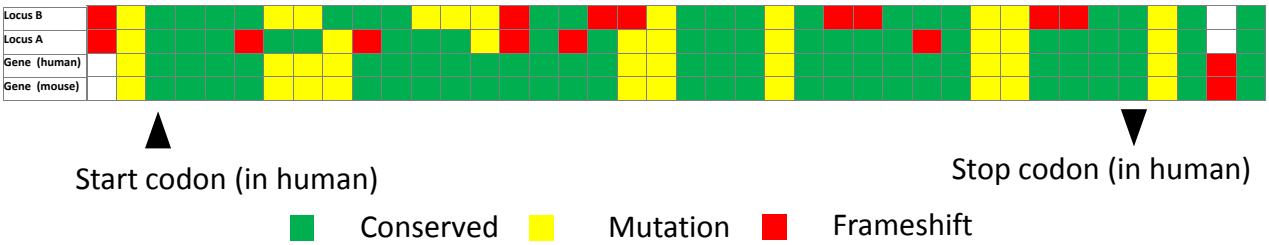
regions that aligned with no gaps in one sequence vs the other.

How many likely exons and introns are there in each of the putative duplicate genes in each of locus 1 and 2?

Locus 1: 3 exons (gray sections) and 2 introns (space between gray sections)

Locus 2: 1 exon and no introns

g) You wish to assess if the duplicate copies are functional or not. For this you align the original gene (Gene (human)) with its most similar gene in mouse (Gene (mouse)) and the putative cDNAs from



each of loci 1 or 2. Which if any of the two loci likely encodes a functional protein? Explain.
Most likely <u>neither</u> locus 1 or 2 encodes a functional protein. Compared to the human and mouse genes, both locus 1 and 2 contain large amounts of frameshifts (red boxes). This would likely lead to a non-functional peptide in both cases.

h) Loci that encode fragmented ORF-like sequences but that no longer encode functional proteins (usually due to many stop codons) are often called 'pseudogenes'. Pseudogenes are either genes that used to be functional and became inactive say with a mutation, or more commonly genes that arose by reverse transcription from RNA and the resulting DNA integrated in the genome (these are called processed pseudogenes). Can you conclude which, if any, of locus 1 or 2 encodes a processed pseudogene? Explain your answer.
Locus 2 contains no introns, so it is likely a processed pseudogene. A reverse transcribed mRNA would also contain no introns due to splicing.

3. Rare white-flowered plants occur in populations of a plant species which normally has purple flowers. The **<u>frequency of the white flowered plant</u>** is $7.4X10^{-4}$. White flowers have an average of 143 seeds, whereas purple flowers have 229 seeds, since the pollinating bumblebee prefers purple flowers to white ones. Interested in this phenomenon, you email your professor friend, a plant developmental molecular geneticist, who tells you that their graduate student has just discovered that the white flowered phenotype is caused by a single **<u>dominant</u>** gene.

a) What is the selective disadvantage (s) and the fitness of the white-flowered plants?

Fitness = 143/229 = 0.624 S = 1-Fitness = .376

b) Assuming that the population is in equilibrium, what is the rate of mutation, μ?

$\mu = Sq$

q= freq(A/a)/2 [because q is rare and dominant, can assume freq(a/a) is negligible]

= 3.7 x 10^-4

μ = Sq= .376*3.7 x 10^-4) =1.39 x 10^-4

c) After happily finishing your calculations, your colleague writes you another email, profusely apologizing. It turns out that the gene is in fact **recessive**. Given this assumption, what is the rate of mutation?

μ = Sq^2

q^2 = freq (a/a) = 7.4 x 10^-4

q= .027

μ = Sq^2= .376*(7.4 x 10^-4) = 2.78 x 10^-4

d) The next year, when you return to the site, you learn that a new pest has emerged in the area where your plants grow. This pest apparently finds that heterozygous flowers are far less tasty (and hence eats them less). Assuming that the heterozygous plants have an advantage h=0.1, What would be the frequency of the white-flowered **allele** (q) after one generation?

Δq sel = -Sq^2 + hq = -.376*(0.027)^2 + (.1)*(.027) = 0.0024

This represents the change after one generation. Must add it to q to get the new value.

q new = 0.027 + .0024 = 0.0294

e) What would be the frequency of **white flowered plants** when the population reaches equilibrium again?

q = h/s = 0.1/0.376 = 0.26

q^2 = .0676 = freq of white flowered plants