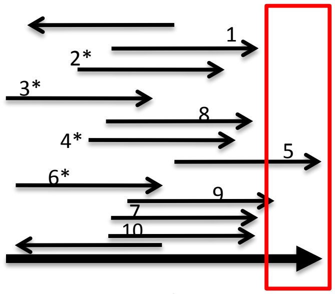
7.03 Problem Set 5

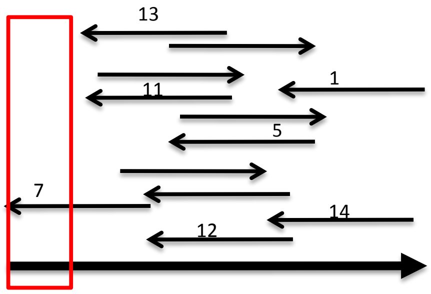
Due before 5 PM on Friday, April 22

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

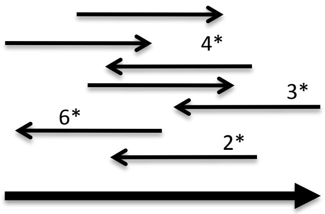
1. You are assembling the genome of a new species, relying on shotgun sequencing. The sequencing was done with two types of clones. One class was made with inserts approximately 2-4Kb in size, and another with much longer inserts, about 100Kb in size. By identifying regions of overlap, you have assembled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



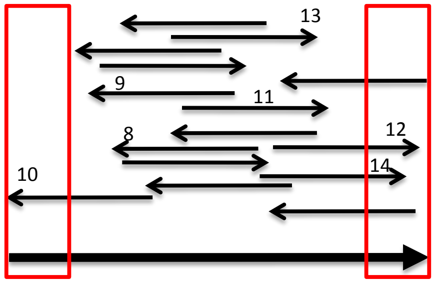
Contig A



Contig C



Contig B



Contig D

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 contig A. Do you agree with the inclusion of the repeat sequence within the contig? Why or why not?

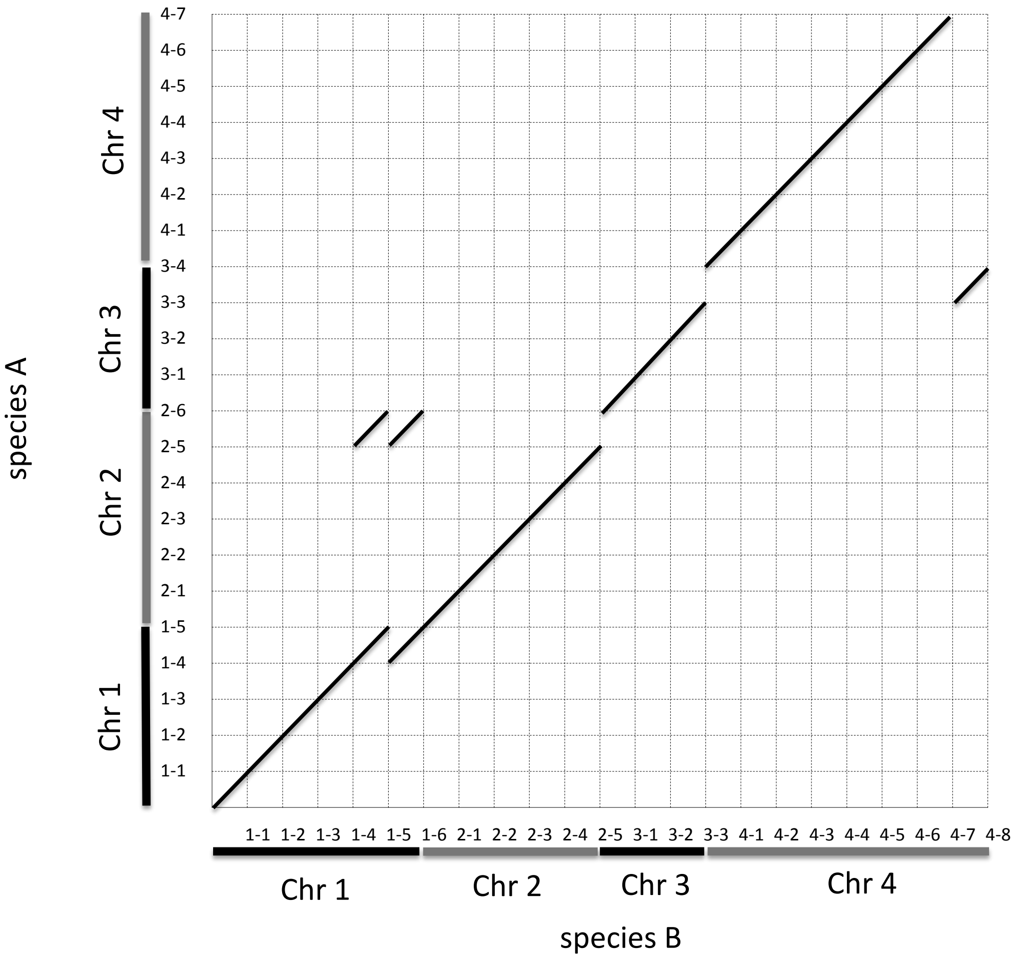
b) Draw the relative organization of contigs A and B. Approximately, how far apart do you expect their sequences to be in the genome? Explain your answer.

c) Now consider contig 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?

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?

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).

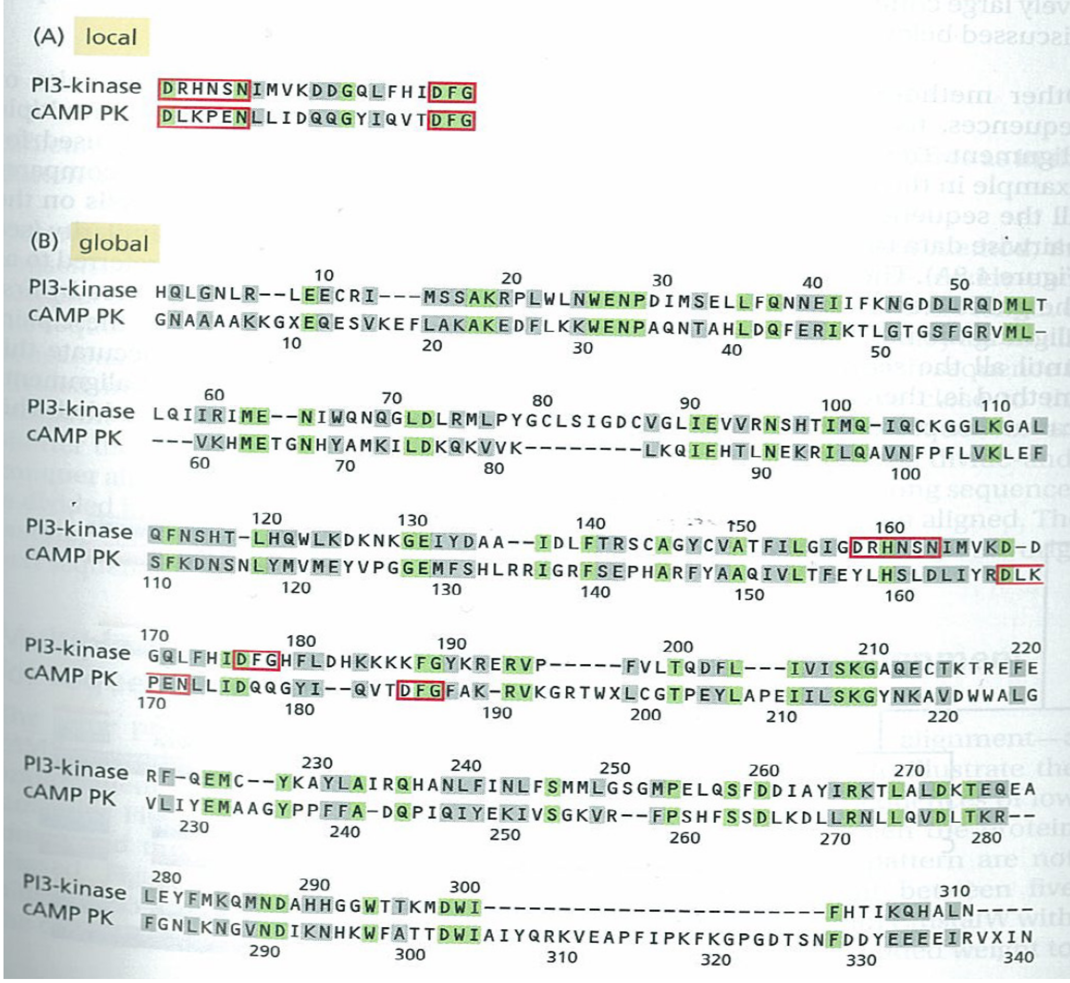
Is this gene present in species B? If so, in how many copies and in which coordinates?



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

g) What happened in species B to the sequence at coordinates 3-3 to 3-4 of species A?

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) local

PI3-kinase  DRHNSNIMVKDDGQLFHIDFG
cAMP PK     DLKPENLLIDQQGYIQVTDFG

(B) global
                 10          20            30          40         50
PI3-kinase  HQLGNLR--LEECRI---MSSAKRPLWLNWENPDIMSELLFQNNEIIFKNGDDLRQDMLT
cAMP PK     GNAAAAKKGXEQESVKEFLAKAKEDFLKKWENPAQNTAHLDQFERIKTLGTGSFGRVML-
                     10        20          30          40         50

               60         70          80          90         100       110
PI3-kinase  LQIIRIME--NIWQNQGLDLRMLPYGCLSIGDCVGLIEVVRNSHTIMQ-IQCKGGLKGAL
cAMP PK     ---VKHMETGNHYAMKILDKQKVVK---------LKQIEHTLNEKRILQAVNFPFLVKLEF
              60        70         80                90        100

                 120        130          140          150        160
PI3-kinase  QFNSHT-LHQWLKDKNKGEIYDAA--IDLFTRSCAGYCVATFILGIGDRHNSNIMVKD-D
cAMP PK     SFKDNSNLYMVMEYVPGGEMFSHLRRIGRFSEPHARFYAAQIVLTFEYLHSLDLIYRDLK
            110       120         130         140          150        160

            170        180         190           200           210       220
PI3-kinase  GQLFHIDFGHFLDHKKKKFGYKRERVP-----FVLTQDFL---IVISKGAQECTKTREFE
cAMP PK     PENLLIDQQGYI--QVTDFGFAK-RVKGRTWXLCGTPEYLAPEIILSKGYNKAVDWWALG
            170       180          190           200         210        220

                     230         240         250          260          270
PI3-kinase  RF-QEMC--YKAYLAIRQHANLFINLFSMMLGSGMPELQSFDDIAYIRKTLALDKTEQEA
cAMP PK     VLIYEMAAGYPPFFA-DQPIQIYEKIVSGKVR--FPSHFSSDLKDLLRNLLQVDLTKR--
              230        240          250           260         270       280

             280       290        300                              310
PI3-kinase  LEYFMKQMNDAHHGGWTTKMDWI------------------------FHTIKQHALN----
cAMP PK     FGNLKNGVNDIKNHKWFATTDWIAIYQRKVEAPFIPKFKGPGDTSNFDDYEEEEIRVXIN
                  290         300         310          320         330        340
```

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

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

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:

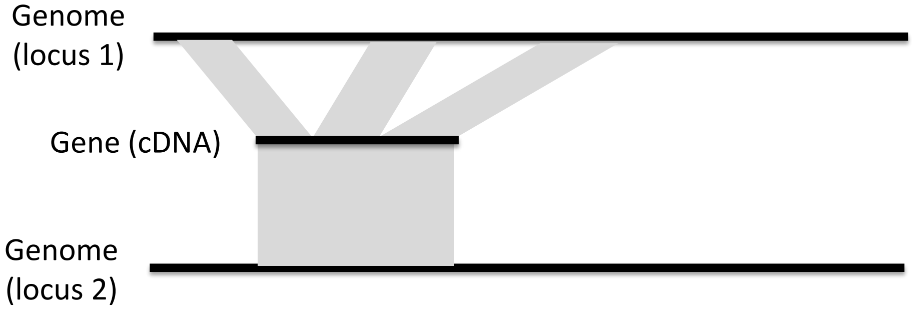
Match: +2
Mismatch: -2
Gap: -8

(Note: we align here two amino acid sequences, not nucleotides, but the procedure is the same).

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

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.

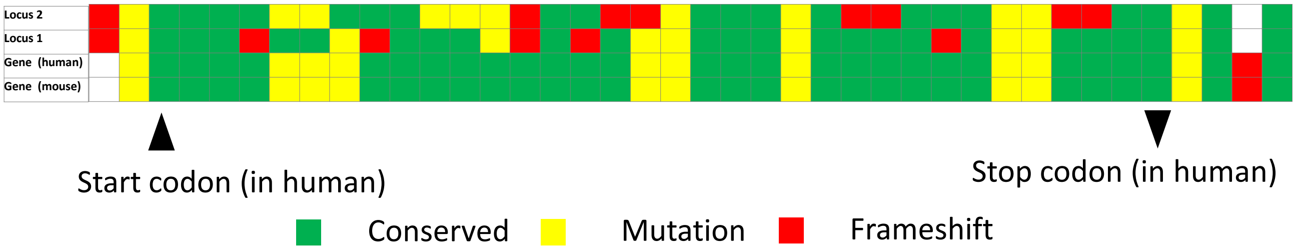
f) By comparing your gene's sequence to the rest of the genome you find significantly similar alignments to two loci (besides the gene locus), which appear below. The grey shading indicates the



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?

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. Does any of the two loci encode a functional protein? If yes, which one and why? If not, why not?



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). Looking at the diagram in (**f**), can you conclude which, if any, of locus 1 or 2 encodes a processed pseudogene? Explain your answer.

3. Rare white-flowered plants occur in populations of a plant species which normally has purple flowers. The **frequency of the white flowered plant** 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 **dominant** gene.

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

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

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?

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?

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