Biol 591 Introduction to Bioinformatics (Fall 2002): Problem Set 1M

P1.1. Most transcriptional factors bind to DNA the same way as CRP. Two of the following sequences are binding sites of actual transcriptional factors. Which do you think they are?

a. `AATTGTGAGCGGATAACAATT`

b. `CACTAGACGGCTGTGATAGT`

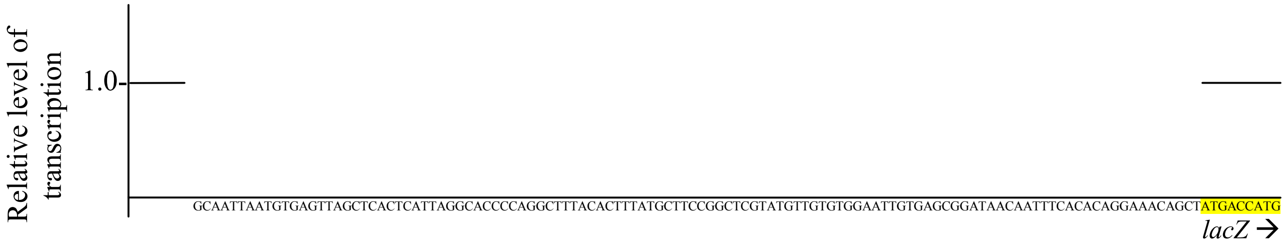
c. `AAATTAGAGACTTGTGACATG`

d. `ACCTGTAGGATCGTACAGG`

P1.2. The gene *lacI*, encoding the Lac repressor, is right next to the *lac* operon. Do you think regulation would be affected if *lacI* were distant from the *lac* operon? Do you think regulation would be affected if the *lac* operator were distant from the *lac* operon?

P1.3. *E.coli* is grown on lactose alone and then switched to a medium containing glucose but no lactose. ß-galactosidase activity is, of course, high prior to the switch, but it remains high for several hours afterwards. Why?

P1.4. The chart below shows on its X-axis the DNA sequence from *E. coli* preceding the *lacZ* gene, from 66 bases prior to the start of *lacZ* to 9 bases within *lacZ* (the same sequence is given below the chart in a larger font for those without microscopic vision). Draw two curves relating the predicted level of transcription of *lacZ* in different mutants. The first curve should reflect transcription when *E. coli* is growing in the presence of lactose and absence of glucose. For the second curve, consider *E.* coli to be growing in the presence of glucose and absence of lactose. The height of the curve at each point should indicate the level of transcription you predict if the nucleotide at that position were deleted, relative to the level of *lacZ* transcription if the nucleotide were not deleted. For example, since you wouldn't think that changing the nucleotides within *lacZ* should affect the transcription of the gene, the height of the far right of the curve should be 1.0.



```
GCAATTAATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGC
TCGTATGTTGTGTGGAATTGTGAGCGGATAACAATTTCACACAGGAAACAGCTATGACCATG
```

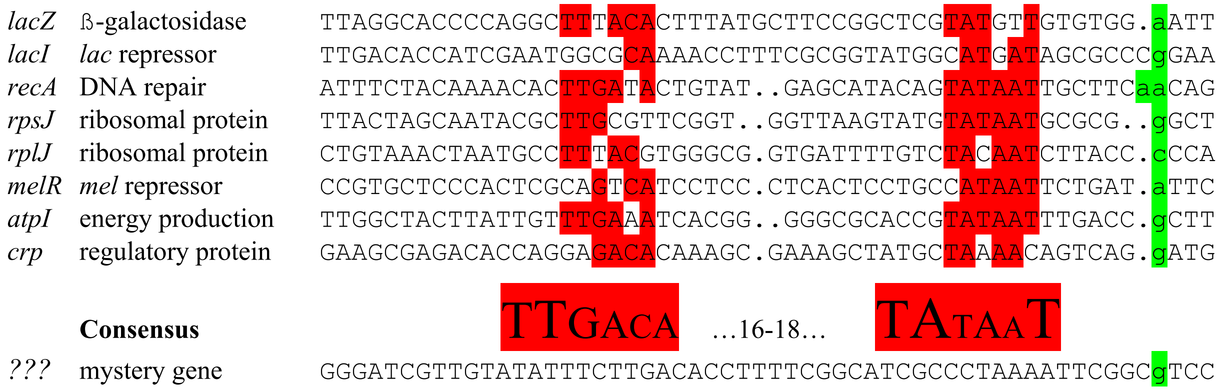
P1.5. A cell in the heart, a cell in the bicep muscle, and a cell in the pancreas have the same set of genes. All three cells express some of the same genes, the heart cell and bicep muscle cell express some genes that the pancreas cell does not, and each cell type expresses genes not expressed in the other cell types. Explain how all this could happen.

P1.6. Just as the *lacZ* gene is preceded by a promoter, i.e. a binding site for RNA polymerase, so are other genes. A list of sequences upstream from some genes is given below. Notice that some genes have sequences that are very close to the consensus promoter sequence while others don't. Ribosomal proteins are amongst the most abundant in a cell (i.e. they are highly expressed), and it so happens that the genes *rpsJ* and *rplJ*, encoding two of the many proteins that make up ribosomes, have promoters that fit closely with the consensus promoter sequence, particularly in the most conserved regions (TT…. and TA…T). On the other hand, two proteins you now know, the lac repressor and CRP, are encoded by genes with promoters that fit poorly with the consensus sequence.

- **a.** Make a prediction as to how strongly *atpI* is expressed.

- **b.** *recA* encodes a protein involved in repair of DNA. The RecA protein binds to DNA looking for damaged regions. The *lac* repressor, the *mel* repressor (which acts analogously, binding near the promoter of the *melA* gene), and the cAMP regulator protein encoded by *crp* all bind to DNA. Rationalize why the genes for three proteins have promoters that poorly match the consensus but the *recA* promoter is a good match.

- **c.** Consider the DNA sequence upstream from the mystery gene. How strongly do you think it is expressed?

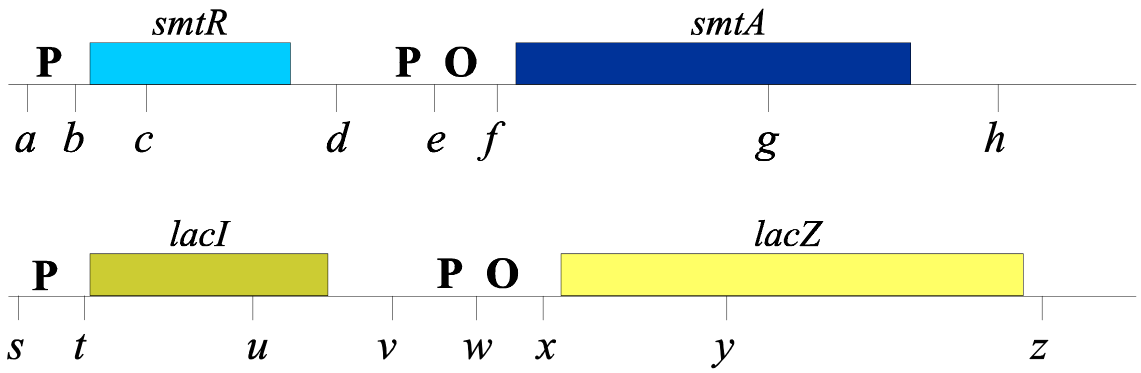


lacZ	ß-galactosidase	TTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTATGTTGTGTGG.aATT
lacI	*lac* repressor	TTGACACCATCGAATGGCGCAAAACCTTTCGCGGTATGGCATGATAGCGCCCgGAA
recA	DNA repair	ATTTCTACAAAACACTTGATACTGTAT..GAGCATACAGTATAATTGCTTCaaCAG
rpsJ	ribosomal protein	TTACTAGCAATACGCTTGCGTTCGGT..GGTTAAGTATGTATAATGCGCG..gGCT
rplJ	ribosomal protein	CTGTAAACTAATGCCTTTACGTGGGCG.GTGATTTTGTCTACAATCTTACC.cCCA
melR	*mel* repressor	CCGTGCTCCCACTCGCAGTCATCCTCC.CTCACTCCTGCCATAATTCTGAT.aTTC
atpI	energy production	TTGGCTACTTATTGTTTGAAATCACGG..GGGCGCACCGTATAATTTGACC.gCTT
crp	regulatory protein	GAAGCGAGACACCAGGAGACACAAAGC.GAAAGCTATGCTAAAACAGTCAG.gATG
	Consensus	TTGACA …16-18… TATAAT
???	mystery gene	GGGATCGTTGTATATTTCTTGACACCTTTTCGGCATCGCCCTAAAATTCGGCgTCC

DNA sequences upstream from several genes. Transcriptional initiation points are shown in green. Regions corresponding to the two binding points for RNA polymerase (comprising the promoter) are shown in red. The degree to which each position of the promoter is conserved (amongst strong promoters) is represented by the height of the letter in the consensus line.

- **d.** (*extra*) Now unmask the gene. Go to the National Center for Biological Information (NCBI; www.ncbi.nlm.nih.gov), click on Blast, click on standard nucleotide-nucleotide Blast, copy the mystery sequence into the box, click the BLAST button, click the FORMAT button (wait), and examine the output to try to find the identity of the gene attached to this sequence.

P1.7. You want to market a home testing unit that will enable households living nearby mining operations to monitor minute levels of toxic heavy metals. Needless to say, your customers will want to know about heavy metals before they reach levels that are lethal. Your plan is to make use of a gene from the bacterium *Synechococcus* PCC 7942, which is highly resistant to heavy metals. *Synechococcus* achieves its resistance by producing in large quantities a protein, metallotheionein, that binds to heavy metals and prevents them from acting on the cell. The protein is encoded by *smtA*. *smtA* is preceded by *smtR*, which encodes a repressor of the gene. You have cloned the *smt* region and from the sequence and other experiments deduced the following map (shown with the previously known map of the *lacZ* region):



Maps of *smtA* and *lacZ* regions. Promoters and operators for the genes are indicated by **P** and **O**, respectively. Lower case letters indicate restriction sites.

Your goal is to cut out part of the *smtA* region and part of the *lacZ* region in such a way so that the two pieces can be ligated together, returned to *Synechococcus*, and used to detect heavy metals. The test you envision is that the customer adds suspect water to a test tube containing the genetically modified *Synechococcus*. Then, the customer adds *o*-nitrophenylgalactoside, a colorless chemical that turns yellow when acted upon by ß-galactosidase. The test is positive or negative depending on the resulting color of the liquid in the tube.

a. Describe exactly what pieces you would use from the *smtA* and *lacZ* regions and how you would put them together to achieve the desired end.

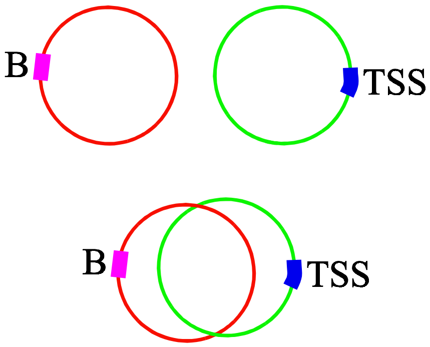
b. (*Extra*) What problems might you encounter in using the organism. For example, what might cause engineered *Synechococcus* to give false positives or negatives?

P1.8. Here's a frequently heard argument: A typical gene consists of about 1000 nucleotides. You'd expect to encounter a sequence of 1000 specific nucleotides with a frequency of one in 4^{1000}, about 1 followed by 300 zeros. Even if the Earth were completely covered a kilometer deep with cells (I calculate about 10^{36} bacterial-sized cells), each churning out 30,000 new gene sequences each second (30,000 being the approximate number of genes in a human cell), you'd need about 10^{252} years before you'd expect the specific gene sequence to arise. This is far more than the billion or so years in which evolution is supposed to have taken place, therefore...

a. What's wrong with this argument?

b. (*Extra*) Do the calculation yourself to get the number 10^{252}. (Hint: I approximated a bacterial cell to be a 1 µm cube).

P1.9. Wedel et al [Science (1990) 248:486-489] studied how postive acting transcription factors work by analyzing the requirements for binding in vitro (in a test tube) of RNA polymerase to the position near the start of transcription. Two situations were examined (shown at right). In the first, a plasmid carrying the binding site (**B**) for a positive acting transcriptional regulator (like CRP) was mixed with a separate plasmid carrying the transcriptional start site (**TSS**) plus 32 bases upstream. In the second, these two plasmids were linked together (like in a chain).



The authors added RNA polymerase, the positive acting transcriptional regulator, plus other necessary components to the two plasmids and measured transcription. They found that transcription was significantly higher when the plasmids were linked than when they were not. What does this result say about what is required for binding sites to enhance transcription?

P1.10. How frequently to certain sequences occur in genomes? One can approach this question through simulation or (if the genome has been sequenced) through an actual count

- **a.** Calculate the probability of encountering the sequence GAGCTC in the genome of *Nostoc*. Do the calculation presuming first that the frequencies of the four bases are all 0.25, then repeat the calculation using the actual frequencies in the *Nostoc* genome for G and C of about 0.2 and for A and T of about 0.3.
- **b.** Calculate the probability of encountering the sequence GACTCG in the same genome, making the same assumptions.
- **c.** Come up with a program that will *count* the number of GAGCTC sequences within the genome of *Nostoc*.
- **d.** Come up with a separate program that will count the number of GACTCG sequences within the same genome.
- **e.** On the basis of these results, assess the validity of using a simulation to predict the frequency of six-base sequences in genomes, where the model takes into account the frequencies of individual nucleotides.