1) Aside from repetitive elements what are some other regions of low complexity that are common in the human genome?

2.) In class we learned that denatured singe-stranded DNA reassociates through defined kinetics, governed by the differential equation:

$$dC/dt = -k * C^2$$

where C is the amount of single-stranded DNA remaining at time t.

a.) Assuming standard metrics, what are the units for k?

b.) Integrate the above differential equation to solve for C / C0, the percentage of single stranded DNA remaining, starting from an initial ssDNA concentration of C = C0. For full credit, show each step of your work.

c.) Given k = 0.5, plot the famous "C0T curve". This curve is drawn on an xy axis, with y = C / C0 and x = C0 * t.

d.) t1/2 is called the "reassociation half time", defined as the time at which exactly half of the DNA has reassociated to become double stranded. Solve for t1/2 in terms of k. Show each step of your work. Label the reassociation half time on your C0T curve from part c.

3) Splicing is the process by which pre-mRNA is modified to remove certain stretches of non-coding sequences called introns; the stretches that remain include protein-coding sequences and are called exons. Sometimes pre-mRNA messages may be spliced in several different ways, allowing a single gene to encode multiple proteins. This process is called alternative splicing.

a.) A gene contains 85 exons. How many different proteins can this system produce?

b.) A gene contains 85 exons of which exactly 24 are to be retained in the final mRNA. How many different proteins can this system produce?

c.) For a gene with 2 exons, what chemical species can be used as evidence that splicing has occurred (other than the final transcript)?

4) Because there are four nucleotides in DNA, adenine (A), guanine (G), cytosine (C) and thymine (T), there are 64 possible triplets encoding 20 amino acids, and three translation termination (nonsense) codons. Because of this degeneracy, all but two amino acids are encoded by more than one triplet. Different organisms often show particular preferences for one of the several codons that encode the same amino acid.

T. maritima is an organism that thrives at 90 degrees C (very hot). How might it choose to code for the peptide “MAGTIDE” if denaturation of the genome is a major concern for survival? Give the DNA sequence and directionality

5) Write an Open Reading Frame (ORF) detector for e-coli.

ORF detectors are a good basic way to detect genes in prokaryotes. We will define an ORF as any sequence that starts with a start codon, ends with a stop codon and does not contain any stop codons in

the same reading frame. For example ATGGTAGGGTAG would contain the codons ATG GTA GGG TAG, the TAG in the middle is ignored because it is in the wrong reading frame. ORFs are only probably real if they are long. Only report ORFs that are over 300 nucleotides long. You will need to search all 6 open reading frames (3 on the + strand and 3 on the – strand)

When reporting the data output this format:

Ecoli start stop strand (either + or -)

For ORFs on the – strand report the location of the bases on the + strand. Meaning the location of the stop codon will the in the start column, and the start codon will be in stop column. Where the start and stop locations are relative to the first base in the fasta file

Practical Information:

Information about e-coli can be bound here:
ftp://ftp.ncbi.nlm.nih.gov/genomes/Bacteria/Escherichia_coli_K_12_substr__DH10B_uid58979/
Download the .fna file, which contains the nucleotide sequence of the genome.

I suggest using a fasta parser to make your life easier. Biopython (www.biopython.org) and Bioperl (www.bioperl.org) both have one. If you feel like it, biojava (www.biojava.org) also has a fasta parser, but I don't recommend using that package.

Programs like this are hard to verify. One of the best ways to verify your results is by comparing against a known list of genes. The .gff file on the nih website contains all known genes in e-coli. You can compare those results to your output file (they will not be exact, but you can at least get an idea if you are in the ballpark). To check your work quickly there is a program called bedtools (http://code.google.com/p/bedtools/) (or pybedtools) that are designed to deal with gff files and your output file (called a bed file). If you are worried about the accuracy of your results use these as a sanity check.

Turn in instructions:

Zip your written answers, your code, and your results and submit it via TED by Midnight Thursday Oct 11th.