CS426: Introduction to Computational Biology

Section 1

TA.

- My name: Liviu Popescu
- email: liviup@cs.cornell.edu
- Office: 424 Rhodes Hall
- Office hours: MW 3:00 – 4:00

Problem 1

- Consider the sequence:

TAATCGAATGGGC

find six possible protein sequence that could derive from it.

Problem 1

	T	C	A	G
T	TTT Phe (F) TTC " TTA Leu (L) TTG "	TCT Ser (S) TCC " TCA " TCG "	TAT Tyr (Y) TAC " TAA **Ter** TAG **Ter**	TGT Cys (C) TGC " TGA **Ter** TGG Trp (W)
C	CTT Leu (L) CTC " CTA " CTG "	CCT Pro (P) CCC " CCA " CCG "	CAT His (H) CAC " CAA Gln (Q) CAG "	CGT Arg (R) CGC " CGA " CGG "
A	ATT Ile (I) ATC " ATA " **ATG** Met (M)	ACT Thr (T) ACC " ACA " ACG "	AAT Asn (N) AAC " AAA Lys (K) AAG "	AGT Ser (S) AGC " AGA Arg (R) AGG "
G	GTT Val (V) GTC " GTA " GTG "	GCT Ala (A) GCC " GCA " GCG "	GAT Asp (D) GAC " GAA Glu (E) GAG "	GGT Gly (G) GGC " GGA " GGG "

S

Problem 1

- Take all possible reading frames
- Reverse the sequences obtained in the first pass

XSNG
NRMG
IEW
GMRN
GNSX
WEI

Problem 2

- Given the fictitious "gene" below find:
  1. the sequence of corresponding mRNA;
  2. the sequence of the resulting protein.

ATGATACCGACGTACGGCATTTAA

TACTATGGCTGCATGCCGTAAATT

Problem 2

- The *coding* strand or the *antisense* strand is the one that looks like mRNA with the exception that all T's are replaced by U.
- The other one is called the *sense*, *anticoding* or *template* strand.
- We consider the first strand to be the *coding* strand and we get the mRNA sequence by replacing T with U

`AUGAUACCGACGUACGGCAUUUAA`

- To get the aminoacid sequence we replace the a codon with the respective aminoacid

`MIPTCGIX`

Problem 3

- Given the protein sequence LMK how many DNA sequences could possibly have given rise to it?

Problem 3

- Each aminoacid is encoded by a number of codons. The number of possible DNA sequences that may encode a certain aminoacid sequence is the product of all the numbers of codons for each aminoacid in the sequence.
- in our case 6*1*2 = 12

Problem 4

- Suppose we have a DNA molecule of length 40000 and digest it with a 4-cutter restriction enzyme. Assuming a random distribution of bases how many pieces can we expect to get?

Restriction enzymes

- proteins with the capacity of cutting the DNA at specific points called restriction sites.
- Restriction sites are specific subsequences for each enzyme. They are palindromes in the sense that each sequence is equal with its reverse complement

Restriction enzymes

- Example *EcoRI* with the cut pattern: GAATTC.
- In this case the the DNA is cut after the first G in both strands. Leaving a number of bases not paired to match cuts from the same enzyme.
- There are several types of restriction enzymes:4-cutters, 6 – cutters, 8 cutters named after the length of their cut pattern.

Problem 4

- The question is how many cut patterns we have.
- We consider all subsequences of length 4 and we get about 40000 subsequences we can consider this as instances in a space defined by 4 random variables. The probability that a certain character would appear is ¼ and of course (¼)^4 for the entire pattern.
- We have 40000 possible cuts and in order to find the expected number of cuts we multiply the possible number of cuts with the probability of a cut.
- We get 40000* (¼)^4 ≈156 cuts and if we consider the number sequences we get 157