7.012 Central Dogma Section

Part 1

Shown below is a 240 base pair segment of a modified version of an *E. coli* gene. It includes the promoter and the first codons of the gene.

The sequences of both strands of the DNA duplex are shown in Figure 1. The top strand reads 5' to 3' left to right (1 to 240); the bottom, complimentary, strand reads 5' to 3' right to left (240 to 1).

```
 5'-ATGTGAGTTAGCTCACTCATTAGGCACCCCAGGCTTTACACTTTATGCTTCCGGCTCGTA
  1 ---------+---------+---------+---------+---------+---------+ 60
 3'-TACACTCAATCGAGTGAGTAATCCGTGGGGTCCGAAATGTGAAATACGAAGGCCGAGCAT


    TGTTGTGTGGAATTGTGAGCGGATAACAATGTCACACAGGAAACAGCTAAGACCATGTTT
 61 ---------+a--------+---------+----d---e+--f-----c+---------+ 120
    ACAACACACCTTAACACTCGCCTATTGTTACAGTGTGTCCTTTGTCGATTCTGGTACAAA


    ACGCCAAGCTCGGAATTAACCCTCACTAAAGGGAACAAAAGCTGGAGCTCCACCGCGGTG
121 ---------+---------+-------g-+---------+---------+---------+ 180
    TGCGGTTCGAGCCTTAATTGGGAGTGATTTCCCTTGTTTTCGACCTCGAGGTGGCGCCAC


    GCGGCCGCTCTAGAACTAGTGGATCCCCCGGGCTGCAGGCATTCGATATCAAGCTTATCG-3'
181 ---------+---------+---------+---------+---------+-x-------+ 240
    CGCCGGCGAGATCTTGATCACCTAGGGGGCCCGACGTCCGTAAGCTATAGTTCGAATAGC-5'
```

a) RNA polymerase binds to the sequence (underlined above) and shown below.

```
5'-...CTTTACACTTT...14bp space....TATGTTG...-3'
     |||||||||||              |||||||
3'-...GAAATGTGAAA...14bp space....ATACAAC...-5'
```

Once bound, RNA polymerase starts making mRNA at the 6th nucleotide after the end of the sequence (at position **a**, also underlined above). Synthesis of the mRNA proceeds 5' to 3' left to right on the sequence above. Write the sequence of the first 10 nucleotides of the resulting mRNA.

b) What are the first five amino acids of the resulting protein?

c) Does translation terminate at the underlined TAA at position 108 (**c**, bold)? Why or why not?

d) How would your answer to b) change if the C/G base pair at position 95 (**d**, bold) was deleted?

e) How would your answer to b) change if an A/T base pair were added between 98 & 99 (**e**, bold)?

f) How would your answer to b) change if the A/T base pair at position 103 (**f**, bold) were changed to G/C?

g) Give a single base change (substitution, deletion, or addition of a single base and it's partner on the other strand) that would cause termination of the polypeptide chain at TAA codon 147 (**g**, underlined).

h) Give a nonsense mutation (codon --> stop codon).

i) Give a missense mutation (codon --> codon for another amino acid).

j) Give a silent mutation (codon ---> codon for the same amino acid).

The Genetic Code

		U	C	A	G	
U		UUU phe	UCU ser	UAU tyr	UGU cys	U
		UUC phe	UCC ser	UAC tyr	UGC cys	C
		UUA leu	UCA ser	**UAA STOP**	**UGA STOP**	A
		UUG leu	UCG ser	**UAG STOP**	UGG trp	G
C		CUU leu	CCU pro	CAU his	CGU arg	U
		CUC leu	CCC pro	CAC his	CGC arg	C
		CUA leu	CCA pro	CAA gln	CGA arg	A
		CUG leu	CCG pro	CAG gln	CGG arg	G
A		AUU ile	ACU thr	AAU asn	AGU ser	U
		AUC ile	ACC thr	AAC asn	AGC ser	C
		AUA ile	ACA thr	AAA lys	AGA arg	A
		AUG met	ACG thr	AAG lys	AGG arg	G
G		GUU val	GCU ala	GAU asp	GGU gly	U
		GUC val	GCC ala	GAC asp	GGC gly	C
		GUA val	GCA ala	GAA glu	GGA gly	A
		GUG val	GCG ala	GAG glu	GGG gly	G

Part 2

Given the sequences on these next two pages, your goal is to draw a schematic of the *con-6* gene. Determine the transcription start and stop sites, start and stop codons, untranslated regions, introns and exons.

```
5'-CGGTGAATAAATACGTCATGACGGTGCTGTCAGCATCATCGATAGGTAGGAGCGAACAAACAACCTAACATCGGATTGCA
  1 +---------+---------+---------+---------+---------+---------+---------+---------+---------
3'-GCCACTTATTTATGCAGTACTGCCACGACAGTCGTAGTAGCTATCCATCCTCGCTTGTTTGTTGGATTGTAGCCTAACGT

    GGACCGCGGGGCAGGATTGCTCCGGGCTGTTTCATGACTTGTCAGGTGGGATGACTTGGATGGAAAAGTAGAAGGTCATG
 81 +---------+---------+---------+---------+---------+---------+---------+---------+---------
    CCTGGCGCCCCGTCCTAACGAGGCCCGACAAAGTACTGAACAGTCCACCCTACTGAACCTACCTTTTCATCTTCCAGTAC

    GGGTGGCCAACTTGGGCGAGAAAAGGTATATAAAGGTCTCTTGCTCCCATCAACTGCCTCAAAAGTAGGTATTCCAGCAG
161 +---------+---------+---------+---------+---------+---------+---------+---------+---------
    CCCACCGGTTGAACCCGCTCTTTTCCATATATTTCCAGAGAACGAGGGTAGTTGACGGAGTTTTCATCCATAAGGTCGTC

    ATCAGACAACCAAACAAACACACTTCATTCCCAAGACATCACTCACAAACAACCAACCTCTTCCAATCCAACCACAAACA
241 +---------+---------+---------+---------+---------+---------+---------+---------+---------
    TAGTCTGTTGGTTTGTTTGTGTGAAGTAAGGGTTCTGTAGTGAGTGTTTGTTGGTTGGAGAAGGTTAGGTTGGTGTTTGT

    AAAATCAGCCAATATGTCCGACTTCGAGAACAAGAACCCCAACAACGTCCTTGGCGGACACAAGGCCACCCTTCACAACC
321 +---------+---------+---------+---------+---------+---------+---------+---------+---------
    TTTTAGTCGGTTATACAGGCTGAAGCTCTTGTTCTTGGGGTTGTTGCAGGAACCGCCTGTGTTCCGGTGGGAAGTGTTGG

    CTAGTATGTATCCTCCTCAGAGCCTCCAGCTTCCGTCCCTCGTCGACATTTCCTTTTTTTTTCATATTACATCCATCCAAG
401 +---------+---------+---------+---------+---------+---------+---------+---------+---------
    GATCATACATAGGAGGAGTCTCGGAGGTCGAAGGCAGGGAGCAGCTGTAAAGGAAAAAAAAGTATAATGTAGGTAGGTTC

    TCCCACAATCCATGACTAACCAGAAATATCACAGATGTTTCCGAGGAAGCCAAGGAGCACTCCAAGAAGGTGCTTGAAAA
481 +---------+---------+---------+---------+---------+---------+---------+---------+---------
    AGGGTGTTAGGTACTGATTGGTCTTTATAGTGTCTACAAAGGCTCCTTCGGTTCCTCGTGAGGTTCTTCCACGAACTTTT

    CGCCGGCGAGGCCTACGATGAGTCTTCTTCGGGCAAGACCACCACCGACGACGGCGACAAGAACCCCGGAAACGTTGCGG
561 +---------+---------+---------+---------+---------+---------+---------+---------+---------
    GCGGCCGCTCCGGATGCTACTCAGAAGAAGCCCGTTCTGGTGGTGGCTGCTGCCGCTGTTCTTGGGGCCTTTGCAACGCC

    GAGGATACAAGGCCACCCTCAACAACCCCAAAGTGTCCGACGAGGCCAAGGAGCACGCCAAGAAGAAGCTTGACGGCCTC
641 +---------+---------+---------+---------+---------+---------+---------+---------+---------
    CTCCTATGTTCCGGTGGGAGTTGTTGGGGTTTCACAGGCTGCTCCGGTTCCTCGTGCGGTTCTTCTTCGAACTGCCGGAG

    GAGTAAGCTCAGAGTTCACGAAAGAACCATTCGACGAGGGGAAGCACGGGGTTATCTCGTTCGAAACATGGGCCTGGTTA
721 +---------+---------+---------+---------+---------+---------+---------+---------+---------
    CTCATTCGAGTCTCAAGTGCTTTCTTGGTAAGCTGCTCCCCTTCGTGCCCCAATAGAGCAAGCTTTGTACCCGGACCAAT

    ATGCAAATGCATAATGGGGAGGATAATGAATCATGAGGTGTACGATATGGACGATATTGACGGATCTTAATTTGATGACA
801 +---------+---------+---------+---------+---------+---------+---------+---------+---------
    TACGTTTACGTATTACCCCTCCTATTACTTAGTACTCCACATGCTATACCTGCTATAACTGCCTAGAATTAAACTACTGT

    GTAATGAAATCACACCATAGT-3'
881 +---------+---------+ 901
    CATTACTTTAGTGTGGTATCA-5'
```
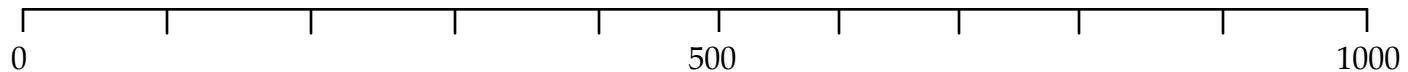
Figure 1: Genomic DNA sequence of *con-6* gene from *Neurospora crassa*. The sequence of both strands (5' to 3' on top, 3' to 5' on bottom) is shown above with nucleotides numbered 1 to 901. The dashed lines are interrupted every tenth nucleotide with a "+".

```
GENOMIC DNA:       251 CAAACAAACACACTTCATTCCCAAGACATCACTCACAAACAACCAACCTC 300
                       |||||||||||||||||||||||||||||||||||||||||||||||||
      mRNA:          1 @AAACAAACACACUUCAUUCCCAAGACAUCACUCACAAACAACCAACCUC 49

GENOMIC DNA:       301 TTCCAATCCAACCACAAACAAAAATCAGCCAATATGTCCGACTTCGAGAA 350
                       |||||||||||||||||||||||||||||||||||||||||||||||||
      mRNA:         50 UUCCAAUCCAACCACAAACAAAAAUCAGCCAAUAUGUCCGACUUCGAGAA 99

GENOMIC DNA:       351 CAAGAACCCCAACAACGTCCTTGGCGGACACAAGGCCACCCTTCACAACC 400
                       |||||||||||||||||||||||||||||||||||||||||||||||||
      mRNA:        100 CAAGAACCCCAACAACGUCCUUGGCGGACACAAGGCCACCCUUCACAACC 149

GENOMIC DNA:       401 CTAGTATGTATCCTCCTCAGAGCCTCCAGCTTCCGTCCCTCGTCGACATT 450
                       |||
      mRNA:        150 CUA............................................... 152

GENOMIC DNA:       451 TCCTTTTTTTTTCATATTACATCCATCCAAGTCCCACAATCCATGACTAAC 500

      mRNA:            ..................................................

GENOMIC DNA:       501 CAGAAATATCACAGATGTTTCCGAGGAAGCCAAGGAGCACTCCAAGAAGG 550
                                    |||||||||||||||||||||||||||||||||||||
      mRNA:        153 ..............AUGUUUCCGAGGAAGCCAAGGAGCACUCCAAGAAGG 188

GENOMIC DNA:       551 TGCTTGAAAACGCCGGCGAGGCCTACGATGAGTCTTCTTCGGGCAAGACC 600
                       |||||||||||||||||||||||||||||||||||||||||||||||||
      mRNA:        189 UGCUUGAAAACGCCGGCGAGGCCUACGAUGAGUCUUCUUCGGGCAAGACC 238

GENOMIC DNA:       601 ACCACCGACGACGGCGACAAGAACCCCGGAAACGTTGCGGGAGGATACAA 650
                       |||||||||||||||||||||||||||||||||||||||||||||||||
      mRNA:        240 ACCACCGACGACGGCGACAAGAACCCCGGAAACGUUGCGGGAGGAUACAA 288

GENOMIC DNA:       651 GGCCACCCTCAACAACCCCAAAGTGTCCGACGAGGCCAAGGAGCACGCCA 700
                       |||||||||||||||||||||||||||||||||||||||||||||||||
      mRNA:        289 GGCCACCCUCAACAACCCCAAAGUGUCCGACGAGGCCAAGGAGCACGCCA 338

GENOMIC DNA:       701 AGAAGAAGCTTGACGGCCTCGAGTAAGCTCAGAGTTCACGAAAGAACCAT 750
                       |||||||||||||||||||||||||||||||||||||||||||||||||
      mRNA:        339 AGAAGAAGCUUGACGGCCUCGAGUAAGCUCAGAGUUCACGAAAGAACCAU 388

GENOMIC DNA:       751 TCGACGAGGGGAAGCACGGGGTTATCTCGTTCGAAACATGGGCCTGGTTA 800
                       |||||||||||||||||||||||||||||||||||||||||||||||||
      mRNA:        389 UCGACGAGGGGAAGCACGGGGUUAUCUCGUUCGAAACAUGGGCCUGGUUA 438

GENOMIC DNA:       801 ATGCAAATGCATAATGGGGAGGATAATGAATCATGAGGTGTACGATATGG 850
                       |||||||||||||||||||||||||||||||||||||||||||||||||
      mRNA:        439 AUGCAAAUGCAUAAUGGGGAGGAUAAUGAAUCAUGAGGUGUACGAUAUGG 488

GENOMIC DNA:       851 ACGATATTGACGGATCTTAATTTGATGACAGTAATGAAATCACACCATAG 900
                       ||||||||||||||||||||||||||||
      mRNA:        489 ACGAUAUUGACGGAUCUUAAUUUGAAAAAAAAAAAAAAAAAAAAAAAAAA 538
```

Figure 2: Sequence alignment of *con-6* genomic DNA and mRNA sequences. The top line of each pair of sequences is the sequence of con-6 genomic DNA. The genomic DNA nucleotides are numbered as in figure 1. The bottom line is the sequence of a *con-6* mRNA isolated from *Neurospora crassa*. The nucleotide numbers of the mRNA begin at the 5' end with #1, and end with #539 at the 3'end. Vertical dashes indicate nucleotides identical in both sequences. Dots indicate nucleotides in the genomic sequence that are not found in the mRNA sequence. (@ represents 5' G-cap)

Part 2 continued

Given the previous figures draw a schematic of the *con-6* gene below. Include the transcription start and stop sites, start and stop codons, untranslated regions, introns and exons.

```
├────┬────┬────┬────┬────┬────┬────┬────┬────┬────┤
0                           500                        1000
```

The Genetic Code

	U	C	A	G	
U	UUU phe UUC phe UUA leu UUG leu	UCU ser UCC ser UCA ser UCG ser	UAU tyr UAC tyr UAA STOP UAG STOP	UGU cys UGC cys UGA STOP UGG trp	U C A G
C	CUU leu CUC leu CUA leu CUG leu	CCU pro CCC pro CCA pro CCG pro	CAU his CAC his CAA gln CAG gln	CGU arg CGC arg CGA arg CGG arg	U C A G
A	AUU ile AUC ile AUA ile AUG met	ACU thr ACC thr ACA thr ACG thr	AAU asn AAC asn AAA lys AAG lys	AGU ser AGC ser AGA arg AGG arg	U C A G
G	GUU val GUC val GUA val GUG val	GCU ala GCC ala GCA ala GCG ala	GAU asp GAC asp GAA glu GAG glu	GGU gly GGC gly GGA gly GGG gly	U C A G