Biology 638/738

Assignment 4

Due 3/15/13, 5PM

1 (8pts). There are two genbank format files (1.gb and 2.gb). Write a script that gives a report including the Locus, Version GI number, last name of the first author on the publication, positions of the CDS, full nucleotide sequence, and the given protein sequence.

For 1.gb, the report should be as follows (note the lack of space characters and numbers in the nucleotides within the nucleotide and protein sequences)

NM\_001183380

296147880

Dujon

1..1134

atgcctcactcagttacaccatccatagaacaagattcgttaaaaattgccattttaggtgctgccggtggtatcgggcagtcgttatcgctgcttttgaaagctcagttgcaataccagttaaaggagagcaaccggagcgttacccacattcatctggctctttacgatgtcaaccaagaagccatcaacggtgttaccgccgacttgtctcatatagacacccccatttccgtgtcgagccactctcctgcaggtggcattgagaactgtttgcataacgcttctattgttgtcattcctgcaggtgttccaagaaaacctggcatgactcgtgatgacttatttaacgtgaatgctggtatcattagccagctcggtgattctattgcagaatgttgtgatctttccaaggtcttcgttcttgtcatttccaaccctgttaattctttagtcccagtgatggtttctaacattcttaagaaccatcctcagtctagaaattccggcattgaaagaaggatcatgggtgtcaccaagctcgacattgtcagagcgtccacttttctacgtgagataaacattgagtcagggctaactcctcgtgttaactccatgcctgacgtccctgtaattggcgggcattctggcgagactattattccgttgttttcacagtcaaacttcctatcgagattaaatgaggatcaattgaaatatttaatacatagagtccaatacggtggtgatgaagtggtcaaggccaagaacggtaaaggtagtgctaccttatcgatggcccatgccggttataagtgtgttgtccaatttgtttctttgttattgggtaacattgagcagatccatggaacctactatgtgccattaaaagatgcgaacaacttccccattgctcctggggcagatcaattattgcctctggtggacggtgcagactactttgccataccattaactattactacaaagggtgtttcctatgtggattatgacatcgttaataggatgaacgacatggaacgcaaccaaatgttgccaatttgcgtctcccagttaaagaaaaatatcgataagggcttggaattcgttgcatcgagatctgcatcatcttaa

MPHSVTPSIEQDSLKIAILGAAGGIGQSLSLLLKAQLQYQLKESNRSVTHIHLALYDVNQEAINGVTADLSHIDTPISVSSHSPAGGIENCLHNASIVVIPAGVPRKPGMTRDDLFNVNAGIISQLGDSIAECCDLSKVFVLVISNPVNSLVPVMVSNILKNHPQSRNSGIERRIMGVTKLDIVRASTFLREINIESGLTPRVNSMPDVPVIGGHSGETIIPLFSQSNFLSRLNEDQLKYLIHRVQYGGDEVVKAKNGKGSATLSMAHAGYKCVVQFVSLLLGNIEQIHGTYYVPLKDANNFPIAPGADQLLPLVDGADYFAIPLTITTKGVSYVDYDIVNRMNDMERNQMLPICVSQLKKNIDKGLEFVASRSASS

2&3 Look at the file chrI.gff. This contains the positions of the genes on chromosome I of the *S. cerevisiae* genome (the chromosome sequence is in chrI.fasta). We're going to focus on the 'CDS' lines (CDS means protein-coding sequences). Note that each gene in the file only has one exon: therefore it only has one CDS line.

2 (6 pts). Write a script that goes through chrI.gff and gets the name and protein id for each CDS line, and prints it out as follows:

Gene YAL069W encodes protein 1 from positions 335 through 649 on the + strand.

Gene YAL068W-A encodes protein 2 on positions 538 through 792 on the + strand.

Gene YAL068C encodes protein 3 on positions 1807 through 2169 on the - strand.

3 (6 pts). Write a script that gets the sequence from the chrI.fasta file and the information from the gff file to print out a fasta file with the coding sequences for each gene. This should look something like this:

>YAL012W

ATGACTCTACAAGAATCTGATAAATTTGCTACCAAGGCCATTCATGCCGGTGAACATGTGGACGTTCACGGTTCCGTGATCGAACCCATTTCTTTGTCCACCACTTTCAAACAATCTTCTCCAGCTAACCCTATCGGTACTTACGAATACTCCAGATCTCAAAATCCTAACAGAGAGAACTTGGAAAGAGCAGTTGCCGCTTTAGAGAACGCTCAATACGGGTTGGCTTTCTCCTCTGGTTCTGCCACCACCGCCACAATCTTGCAATCGCTTCCTCAGGGCTCCCATGCGGTCTCTATCGGTGATGTGTACGGTGGTACCCACAGATACTTCACCAAAGTCGCCAACGCTCACGGTGTGGAAACCTCCTTCACTAACGATTTGTTGAACGATCTACCTCAATTGATAAAGGAAAACACCAAATTGGTCTGGATCGAAACCCCAACCAACCCAACTTTGAAGGTCACCGACATCCAAAAGGTGGCAGACCTTATCAAGAAGCACGCTGCCGGCCAAGACGTGATCTTGGTTGTCGACAACACCTTCTTGTCCCCATATATCTCCAATCCATTGAACTTCGGTGCAGACATCGTTGTCCACTCCGCTACAAAGTACATCAACGGTCACTCAGACGTTGTGCTCGGTGTCCTGGCCACTAATAACAAGCCATTGTACGAGCGTCTGCAGTTCTTACAAAACGCCATTGGTGCTATCCCATCTCCTTTCGATGCTTGGTTGACCCACAGAGGTTTGAAGACTTTGCATCTACGTGTCAGACAAGCTGCCCTCAGCGCCAACAAAATCGCTGAATTCTTGGCAGCAGACAAGGAAAACGTTGTCGCAGTCAACTACCCAGGTTTGAAGACACACCCTAACTACGACGTAGTGTTAAAGCAACACCGTGATGCCCTTGGTGGTGGTATGATCTCCTTCAGAATCAAGGGTGGTGCTGAAGCTGCTTCCAAGTTCGCCTCCTCCACAAGACTGTTCACATTGGCCGAATCCCTTGGTGGTATCGAATCTCTATTGGAAGTGCCCGCTGTGATGACCCACGGTGGTATCCCAAAGGAGGCCAGAGAGGCCTCTGGTGTTTTTGACGACTTGGTTAGAATCTCTGTCGGTATTGAAGACACTGACGATCTTTTGGAAGACATCAAGCAAGCCTTGAAACAAGCCACCAACTAA

>YAL047C

TTAGGGATTGTTGATTGATAGGTTGAAAAGTTTCGATCTCAATGACTCATTTTCCTGTTCTAAAGCCTTGATTCGAGCTTCTGAAGCGTTAGCATCGAGCTTCCGTCTTTCTCTTTCGGATATCCATCTTCTTTGTAACTCCTCTATTCGTAAGGTCAGTTCTTTATTTGGAGTTGCCGTTAATTCATTGCCCTGTTGATGTGGCTGCTCTGGGGTTTCCATGGAAATCAGTGAAGAGATATATGAGTTTATTATGGACTCTAATGCAGTCTCTATAAAAGTGTAGAGCGATTCTAGTTTGGGCTGAATCAAATTCAAGTTTTTCAACGCATTTGGTACAGATTTTATGGATTTCATTTTCCTGTCAAATTGGGCAATAGAACTTTCTTGTAAGATTTTCTGCAGAATATTGAAAATACTGTCAAGGTGTAGTAAAAGGTTCTGATTAAATTTGATAAAATTGCTTTCATATTGGGATAGTATTAACTTTTGGTTGTCCAAAGTATTTAACTGATTTTTTAATTGGTAATTCTCCAAGTCTAATTTGTCTAACTGATTTTGTATTACGTGGAATTCCTCTGATTTATCGATTTGAAGGTCATTGATTTGTTTTTCCAAGTCATTGATGTAGCTGTCCCAATTATTTTCCTTTAGTTTATTGATTTTGGTCTGAGTTTCCAGTTCTTTGGTTAAAACTTTTTCATTTTGTTTCATTTTAATGATGTCTTCCTTCAATTTTTCCAAATTGTTTATCAAAACAGATTGCGATTCGATCTTTTCCTTCAGTTGGGAAATCAAATGATCTTGTTTTTCTATAACGGAACTAGAATTTTCTTCCAATTCTAAGGAATCTAGTAGATGAGACTGCTCATTTAGTTTGGACGCTATTATTTTTTCTAATTTTTGCGATTTCTCGAATTTCAATCTAATGGAATTTATAAATTGATCATATTCTTTGTGCAAAATTTCTATGACAATCTCTAATTGAGTGTCTAAGGTTTTCTCAAAACGAGACTCGACGGCTGAAATTGGTAAGGAAGTACTGTTTGCCATTACATTTTCTGTATCAATGTGGTTGTTACTGTCATGAATTTCGTCATTCTGGTATCCACCATCAGTATTCTCCTCTTTACTTGATGGTGAGTCCCTACTTTCTAACTGTGATCCTGCTGGTGAGGACTGGGCCAGTGAAAGAAATTCTTCCTTATCCTGCTTTGATTCCGCTCGACTTTTGGAATGCAAAAAGTCCTGCAAGAATTGGATGATGAACTTGGACAAAGTATCCATTTTTTCAAGAACATAATCCGAGCTCAGCTCTAAAGTTTCCTCCAGGGTCTCCCTCTCTTCTTTATCAAGAAGATGGGCATTTTCATCTTGCTCATTAAAATGCGTCAATATAAAGGACAATAGGTGATTGATCGTATTCACGATACTTTCCGAATTTTCTAAATTCTCTTGAACGAATTGTAGCGTATCTTGGTACTCGACTTCCTTCGCCTTTAAATCCTGTTTCAATTTGTTTATTTCAAGATTTAGACCCTCGATAATCGAATTTCTAAAATCAGTGTCATTGCCCAGCGATGGTGCATTGCCGTCTTTATTAGGGATTCTGCGAATGTATTCATAGAGTACTTGAATCTTGATTTTGGCATTAGTCAACTCCTTCTCCAAATTTTTGACTTTGTTGGAATCGTTCATCAGAGCAGGCTTGATGGGATCGTTATGAGATGACCTCGTGGTCATGGAGTCCCTGAGTGATGGTATGGACATCCCAGAATCCATCGAGTTTGTGAACTCGCTGTCGTCATCATCACCAGTGTTGTCATTATTGCGAAGATGCCTGCCACTAGGAATCCATCGACGTACCAT