CSC-314: Final Project Rubric

Option A Projects

Note: ALL PROGRAMS SUBMITTED MUST BE YOUR OWN. IF ANY PORTIONS OF YOUR PROGRAM ARE FOUND TO BE PLAGIARIZED, THE ASSIGNMENT WILL NOT BE ACCEPTED. If you have any questions about the acceptability of your code, or whether the use of packages or libraries is allowed, you should contact me.

For programming projects, documentation is worth 20% and program functionality is the worth the remaining 80%. In addition to your code, appropriate input files must also be submitted.

	Poor (C or below)	Acceptable (B range)	Excellent (A range)
Documentation (all programs)	Code documentation is minimal or not provided. Variable and function names are not descriptive and the code is formatted poorly, making the program difficult to read.	The majority of functions and major code segments are documented. Variable and function names are chosen appropriately and proper formatting (such as indentation) is used in the majority of the code. The algorithm is easy to follow but some aspects of the code are not.	All functions and major code segments are properly documented. Variable and function names are chosen appropriately and proper formatting (such as indentation) is used, making the program logic easy to follow.
Input files (all programs)	No input files are provided		An appropriate input file or files are provided. Your code will be tested on this input file and possibly others. Even if your code is not completely working, the input file provided will be used to award partial credit, if appropriate.
Translation Program	The program does not work correctly (sequences are not translated correctly)	The program works correctly but is not user-friendly. For example, the user cannot specify the name of the file.	The program works correctly. DNA/RNA sequences are read from a specified file (it is assumed that the sequence is written in the 5' to 3' direction). The program translates all 6 reading frames, and open reading frames are highlighted. The program is user-friendly and handles both lower- and upper-case nucleotide characters.
Pairwise Alignment Program	The program does not work correctly (the score of the optimal alignment is not correct and the optimal alignment cannot be found)	The program works correctly but is not user-friendly. For example, the alignment is correct but is not displayed in the standard format.	The program works correctly and is not case-sensitive. The user can upload two sequences stored in a single file or stored in separate files. A dynamic programming matrix is created in order to find the score of the optimal alignment and a traceback procedure is used to find and display the optimal alignment.

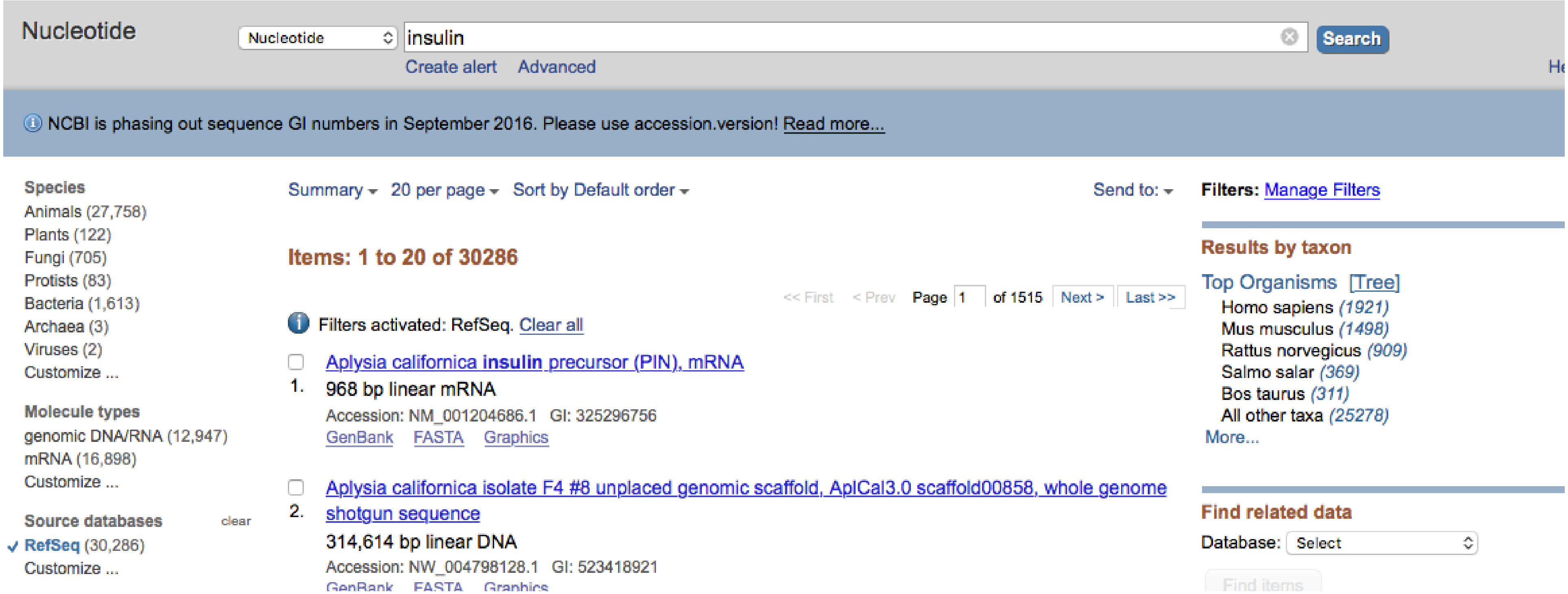
Gene Prediction Program	The program does not work correctly, and either does not identify likely genes, or mistakenly identifies regions that do not have all of the desired properties.	The program has small mistakes that prevent correct identification of all potential genes. For example, the program requires that the Shine-Dalgarno sequence matches exactly, rather than allowing mismatches.	The program is user friendly, works correctly, and is not case-sensitive. The user uploads a file containing one or more sequences in FASTA format. For each sequence, the program outputs the number of predicted genes, and for each predicted gene, the location and length of the CDS.
Exon/Intron Boundaries	The program does not work correctly, for example does not read in the sequence data and does not sufficiently (attempt) to characterize the exon/intron boundaries	The program has small mistakes that prevent the correct characterization of exon/intron boundaries; or doesn't output the results correctly to the file; or doesn't generate a correct graph of the results	The program correctly reads in sequence records in FASTA format and correctly extracts the introns. For all introns, the program correctly counts the number of nucleotides at each position (the first 4 and last 4 nucleotides of an intron), and outputs the results to a csv file. The csv file is converted to an Excel spreadsheet and a stacked bar graph is produced that summarizes the consensus sequences for the exon/intron boundaries. The correct input file or files are obtained from the UCSC table browser.
~~Viterbi Algorithm~~	~~The program does not work correctly.~~	~~The program has minor mistakes and that output is not correct. For example, the correct probabilities are calculated, but the traceback is not correct, and the correct optimal state is not displayed.~~	~~The program is user friendly, and works correctly, generating the optimal sequence of hidden states for 3 coin tosses, where observed values (e.g., *heads*, *tails*, *heads*) are specified by the user. Probabilities are on the *log2* scale. The HMM follows the example we did in class.~~
~~BLAST~~	~~The program does not work correctly.~~	~~The program works correctly but is not complete. For example, the organism and E-values are not correct.~~	~~The program works correctly, BLASTING the sequences using the appropriate parameters and for the top two hits of each BLAST, outputs the organism, accession number, identity, similarity, e-value, and alignment of the top hit for each sequence. The program is user-friendly and allows the user to enter one or two protein IDs. No values in your program should be hardcoded (i.e., if I run the program with another sequence, I should get the correct results for that sequence).~~

Option B Projects

	Poor (C or below)	Acceptable (B range)	Excellent (A range)
Questions (25%)	The majority of your questions have not been answered or are not correct	A small number of questions have not been answered or are not correct.	All or nearly all questions are answered correctly.
Write-up (25%)	Answers are not submitted in the form of a written report that forms a cohesive narrative.	The write-up is a cohesive narrative, but may include several spelling or grammatical mistakes.	The write-up is a cohesive narrative. Little or no spelling or grammatical mistakes are made.
Methods (25%)	No methods are included.	Almost all methods are included.	The write up describes all methods used to obtain your answers. These methods should be specific enough to be repeatable (i.e., so I can get the same answers as you; see below). You may choose to include a separate Methods section or integrate the methods with your answers (see below).
Screenshots (25%)	No screenshots are included	Almost all screenshots are included	Screen shots of all key results and methods are included, which includes but is not limited to: GenBank queries, GenBank or GenPept entries, OMIM entries, BLAST results~~, and GEO2R results.~~

Example question: How many RefSeq entries are there for the keyword "insulin"?

Example answer: There are 30,286 RefSeq entries for the keyword "insulin". This was determined by searching for "insulin" (without the quotes) from the GenBank website (https://www.ncbi.nlm.nih.gov/genbank/). From the side panel, selecting RefSeq under source databases yielded 30,286 results (see screenshot on next page).



Nucleotide

Nucleotide | insulin | Search

Create alert   Advanced

NCBI is phasing out sequence GI numbers in September 2016. Please use accession.version! Read more...

Species
Animals (27,758)
Plants (122)
Fungi (705)
Protists (83)
Bacteria (1,613)
Archaea (3)
Viruses (2)
Customize ...

Molecule types
genomic DNA/RNA (12,947)
mRNA (16,898)
Customize ...

Source databases   clear
✓ RefSeq (30,286)
Customize ...

Summary ▾ 20 per page ▾ Sort by Default order ▾   Send to: ▾

Items: 1 to 20 of 30286

<< First < Prev Page 1 of 1515 Next > Last >>

Filters activated: RefSeq. Clear all

□ 1. Aplysia californica insulin precursor (PIN), mRNA
968 bp linear mRNA
Accession: NM_001204686.1   GI: 325296756
GenBank   FASTA   Graphics

□ 2. Aplysia californica isolate F4 #8 unplaced genomic scaffold, AplCal3.0 scaffold00858, whole genome shotgun sequence
314,614 bp linear DNA
Accession: NW_004798128.1   GI: 523418921
GenBank   FASTA   Graphics

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]
Homo sapiens *(1921)*
Mus musculus *(1498)*
Rattus norvegicus *(909)*
Salmo salar *(369)*
Bos taurus *(311)*
All other taxa *(25278)*
More...

Find related data

Database: Select

Find items

Note for group projects:

1. For Option A projects, documentation must include the person responsible for writing the code (this will generally be at the function / module level)
2. For Option B projects, you must include a contribution section at the end which describes the contribution for each individual. In this section, it is customary to use initials (e.g., GD instead of Garrett Dancik). *Example:* GD retrieved sequences from GenBank, and performed the BLAST analysis. AFP analyzed the gene expression data using the Gene Expression Omnibus (GEO) and collected relevant information from the OMIM database.