**BT-3172: Special Topics in Bioinformatics: Practical computing for bioinformatics**

**Lab 6: Use of Python modules for bioinformatics.**

**Name: Anushka Udara**

**Index number: s14234**

In this practical, you will learn how to import and use built-in or third-party Python modules and packages to solve biological questions.

For this practical, again, you will be working with several genes and proteins involved with the plant stress response pathways. Here, you will be expanding from ABA-independent (e.g., DREB proteins) to ABA-dependent pathway.

After using PyCharm to write your scripts, **copy the codes to the appropriate space below the questions**. Also, submit the Python file separately so they can be tested.

1. Using Biopython and re modules.
   1. Search the NCBI RefSeq **Gene** database for *Arabidopsis thaliana* DREB2A gene. Write its gene ID below. Briefly explain its functions. Locate its genomic sequence record from the GenBank. Write its accession ID with the version number. What is the length of the gene sequence according to the GenBank record? Download the gene sequence in FASTA format and name the file as “ATdreb2a.fasta”.

ID – 830424

Functions - significant drought stress tolerance but only slight

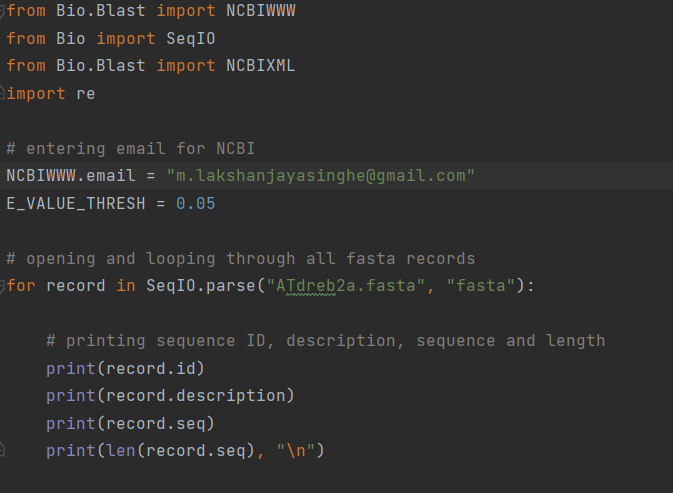
freezing tolerance in transgenic Arabidopsis plants.

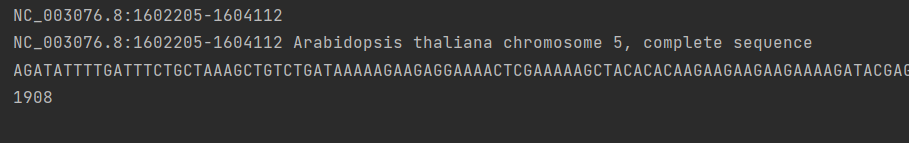
Accession ID - NC\_003076.8

Length - 1908 bp

* 1. Use the Biopython module to perform the following tasks.

Load the downloaded FASTA file as a sequence record object and print the following attributes of the record: sequence ID, description, sequence, and sequence length. Write the answers below.

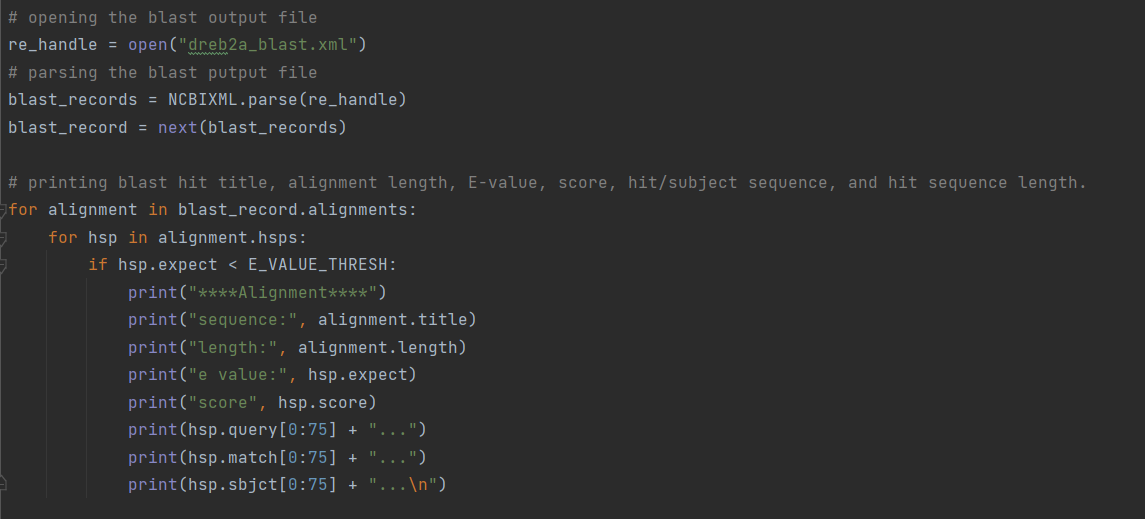


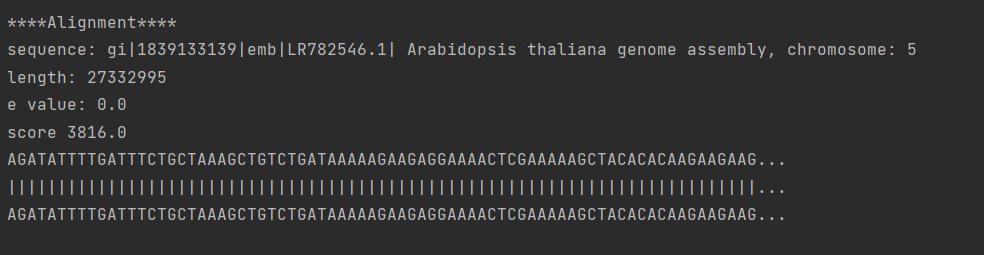


* 1. Using the Biopython module run the web-based nucleotide blast program on the ATDREB2A sequence. Refer to the Biopython manual for instructions. Save the blast output as “dreb2a\_blast.xml”

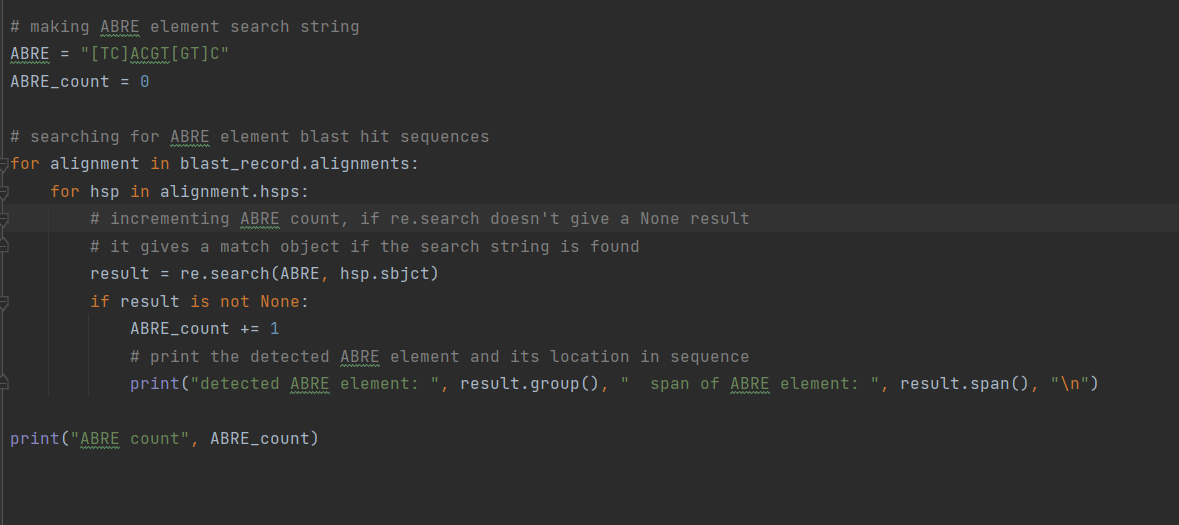


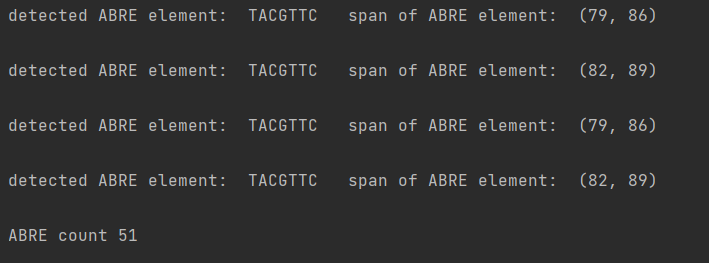
* 1. Parsing the blast output file. Open the xml file saved during the previous question, and parse through the blast hits. Define an E-value threshold of 0.05 and only select the hits that are below the threshold. Print the following attributes of each blast hit selected based on the above criteria: blast hit title, alignment length, E-value, score, hit/subject sequence, and hit sequence length.





* 1. Now, modify the above code to identify the blast hits with the ABRE cis-acting ABA-dependent transcription factor binding element using Python regular expressions. First, define a search string to search for the ABRE element (PyACGTG/TC). Then check each blast hit sequence for the presence of the ABRE element and print the detected sequence fragment (e.g., PyACGTGC or PyACGTTC; “Py” refers to a pyrimidine base) along with the sequence location of each finding. **Please note that the ABRE element can be found in multiple locations in the same sequence**. Further, print the number of blast hits with ABRE element present in the sequence and write that number below.





ABRE count = 51