Iterative Blast README

1. Requirements
    - Local machine:
        i. Biopython 1.77 installed. (For downloading and setting up please visit https://biopython.org/)
        ii. Python 3.6 or greater to support Biopython 1.77
        iii. Command-line ncbi blast. The installers and source code are available from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/.
    - JHPCE Cluster:
        i. Enter “module load python/3.7” and “module load blast” before running the script.
2. Build custom database
    - If needed, you may build a custom database with the reference genomes instead of downloading built-in blast databases from ncbi. The following tutorial is for mac users.
    - Open Applications/Utilities/Terminal
    - Type "mkdir -p ~/blast/db" in Terminal to make a path for the database.
    - Combine all your fasta files into one file called testdb.fasta using the command cat all_file_names > testdb.fasta
    - Copy testdb.fasta to ~/blast/db.
    - Type "cd ~/blast/db" to get to the path
    - Type "makeblastdb -in testdb.fasta -out testdb -dbtype nucl" for DNA or "makeblastdb -in testdb.fasta -out testdb -dbtype prot" for Protein and type return to get the database named testdb.
    - Type "makeblastdb -help" for advanced options.
    - For more info, check https://www.ncbi.nlm.nih.gov/books/NBK279688/
3. Input: a FASTA file containing multiple query sequences
4. Results
    - A txt file named “sequence name + stage name + RESULTS” that records the final annotation results (see the next section for more information.)
    - Fasta files of each individual sequence in the original file before iterations with the sequence name as filename
    - Fasta files named “sequence name + Version #” contain the partially masked sequences after iteration #
    - Fasta files named “sequence name_finished” contain the maximiumly masked sequences after all iterations
    - Log files for each individual sequence in the query file containing a summary of the annotatons.
    - A txt file named either “EXIT_FAILURE” or “EXIT_SUCCESS”

- A txt file containing the title information of all the sequences
- A result.xml file that contains the last blast result

5. How to interpret the final annotation text file:
    - <*** Round # ***>
      > SeqID1 and other text from the title line for this sequence
      -------#####---------111---------22222--------
      1 Query QE SS SE Identities SubjectSeqID SubjectDefinitionLine / No Blast Hits
      2 Query QE SS SE Identities SubjectSeqID SubjectDefinitionLine / No Blast Hits
      …
      # Query QE SS SE Identities SubjectSeqID SubjectDefinitionLine / No Blast Hits

      > SeqID2 and other text from the title line for this sequence
      -------#####---------111---------22222--------
      1 Query QE SS SE Identities SubjectSeqID SubjectDefinitionLine / No Blast Hits
      2 Query QE SS SE Identities SubjectSeqID SubjectDefinitionLine / No Blast Hits
      …
      # Query QE SS SE Identities SubjectSeqID SubjectDefinitionLine / No Blast Hits

      …

      > LastSeqID and other text from the title line for this sequence
      -------#####---------111---------22222--------
      1 Query QE SS SE Identities SubjectSeqID SubjectDefinitionLine / No Blast Hits
      2 Query QE SS SE Identities SubjectSeqID SubjectDefinitionLine / No Blast Hits
      …
      # Query QE SS SE Identities SubjectSeqID SubjectDefinitionLine / No Blast Hits

      <*** Round #+1 ***>
      …

    - The dash is exactly 100 characters long, and the number # shows the proportional region of the query that has the match during the #th iteration
    - QS, QE, SS, SE are the positions of the query start, query end, subject start, subject end. Identities is the Matches/TotalLength result given by Blast. SubjectSeqID and SubjectDefinition line are the information about the subject sequence.

6. Notes
    - The blast outformat is set to be 5 (xml format)
    - There are two cutoffs for a valid match: first an e value cutoff with e=0.1**30, and then an over 95% match between the query and subject sequence