**MATLAB-PLISH H-Probe Design Script: Set-up and Use 2/10/2020**

**Requirements:**

1. MATLAB 2017a (or newer)
2. MATLAB 2017 Bioinformatics Toolbox (or newer)
3. NCBI Blast 2.10+
4. NCBI Refseq-RNA database V5

**Notes:**

1. The script works on a windows system. Though it should work on Mac OS, it will require modifications to the code to support the different file system in Mac OS. Also, as of MATLAB version R2017a, the MATLAB function responsible for saving data an Excel \*.xlsx file is not supported in Macintosh OS. Therefore, only option 1 and 2 (detailed below) of the script will work.
2. A stock MATLAB function included in the R2017a Bioinformatics Toolbox called blastreadlocal was found to contain an error. A modification to the code of blastreadlocal is required to allow it to appropriate parse the blast output file and avoid producing an error. To correct the potential error, open the MATLAB function blastreadlocal in the MATLAB m-file editor. Once open, navigate down to line 492. Line 492 should be changed from:

if strfind(blasttext,**'Strand ='**) % Strand is included with nucleotide sequences

to

if strfind(blasttext,**'Strand='**) % Strand is included with nucleotide sequences

**Installation Process:**

1. This assumes the user already has MATLAB installed. For more information on MATLAB installation, see https://www.mathworks.com/?s\_tid=gn\_logo.
2. First, install NCBI blast on your computer. Navigate to:
   * https://www.ncbi.nlm.nih.gov/books/NBK279671/

for instructions on how to download and install BLAST for various operating systems. The installation files are currently located at

* ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/.

1. The blast in installation instructions describe how to define Widows environmental variables that specify the location of the blast commands (environment variable PATH) and the location of the Refseq database (environmental variable BLASTDB). Defining these variables allows MATLAB and other programs to call the blast program from a command line. For reference, the directories used on our systems and coded into the script are:
   * PATH: C:\NCBI\blast-2.10.0+\bin
   * BLASTDB: C:\NCBI\blast-2.10.0+\db
2. To determine if BLAST has been installed and configured correctly, first type blastn at the DOS prompt in a cmd window (for windows OS). This should generate an error that indicates the blasn command was found and run, but that additional files and parameters need to be specified. To test that BLAST works through MATLAB, type [status,cmdout] = dos(‘blastn’) at the MATLAB prompt. This should return “0” for status and the same blast error in cmdout that was returned in the cmd window. If these steps to not work as expected, then double check the environmental variable step above.
3. Install the NCBI blast reference database. The script uses the RefSeq RNA database which is then filtered on the basis of a taxid provided by the user. To download the most current RefSeq RNA database, navigate to:

https://www.ncbi.nlm.nih.gov/books/NBK62345/#blast\_ftp\_site.The\_blastdb\_subdirectory.

Alternatively, one can use the “update\_blastdb.pl” command, described here:

https://www.ncbi.nlm.nih.gov/books/NBK537770/

If unable to use the update\_blastdb.pl command, directly download the required files from the ftp site here:

ftp://ftp.ncbi.nlm.nih.gov/blast/.

After downloading, unpack and un-tar the downloaded files into the BLASTDB directory identified above.

1. Copy the MATLAB PLISH probe design script (PLISH\_Probe\_Design\_BLAST210.m) to a working directory. Start MATLAB. Set the working directory to the one containing the script.
2. Open the script in the MATLAB m-file editor. The following items may need to be changed:
   * Under the heading called, “% USER SPECIFIED OPTIONS HEADER:”, navigate down to item number 6: “library\_directory”. The identity of the directory should be specified by the user, and it must exist prior to running the MATLAB script. The default is:

'C:\Users\Lab\Desktop\hprobelibrary'.

Each time a new probeset is designed for a specific gene, a new directory of the form ..\<gene name>\_ver\_02\_2020 is created inside the library directory where results from the script are saved.

**Running the script: Option 1**

1. After opening MATLAB and changing MATLAB’s working directory to the directory containing the design script, type PLISH\_Probe\_Design\_BLAST210 at the MATLAB prompt.
2. A menu opens titled, “Select the Mode of Operation'” that allows the user to choose between: 1. Making a probeset from scratch, 2. Opening and re-analyzing a previously created probeset, and 3. Making an excel sheet of desired probes for ordering purposes. For most cases, the user will use option 1 or 3. The following assumes the user chooses option 1.
3. A box opens asking the user to ‘Enter List of Transcript Variant Accession Numbers (IE NM\_001204.6)'. This step is optional but allows for a check to see if a given probe matches all the possible transcript variants for a gene. Enter into the box a list of refseq accession numbers (IE NM\_XXX) corresponding to all the transcript variants of the gene separated by returns. Transcript variants can be identified by searching for the HGNC approved gene symbol found at:

https://www.genenames.org

In the NCBI Nucleotide repository

https://www.ncbi.nlm.nih.gov/nuccore.

1. Next, the program will ask for the nucleotide sequence for the gene of interest. This can be found by searching for the gene in the NCBI Nucleotide repository. At the bottom of the gene’s information page, the sequence will be found immediately after “origin.” Highlight the entire sequence (numbers, whitespace, and all) copy it, and paste it into the open MATLAB box.
2. At the next box, enter the gene name for the sequence you entered. We typically use the HGNC approved gene symbol preceded by ‘h’ for *Homo sapiens* and ‘m’ for *Mus musculus*.
3. The next box asks the user to enter a Bitscore. Based off of experience, we have considered any alignment that generates a bitscore greater than 30 as being a significant alignment and possessing undesirable off-target homology. The script will count up the number of off target alignments as defined by the bitscore and then use this total to rank potential h-probes on the basis of off target homology. The stringency of the script can be adjusted by the user by choosing bitscore values less than 30 (more conservative with respect to off-target homology) or greater than 30 (more liberal). If it is decided to change this stringency later, the script can be re-run by selecting item 2 on the initial prompt. This will allow the bitscore to be changed and the results will be adjusted accordingly. Note option 2 will not repeat any of the blast alignment work which was done in step 1.
4. After choosing the biscore, the program will then ask for a taxid to filter the blast database with. Taxid’s for a given taxon can be found by searching at:

https://www.ncbi.nlm.nih.gov/taxonomy.

For reference, *Homo sapiens* is 9606. *Mus musculus* is 10090. Any taxa contained in the Refseq RNA database can be specified.

1. The program will now parse the nucleotide sequence and select potential h-probe sites. Promising sites are identified by looking for nucleotides which are thought to result in more stable Holliday junctions. Specifically, most probes sites are chosen by identifying 5' – AG - 3' or 5' - TA - 3' sites within the nucleotide sequence. Additionally, sites defined by 5' - ACC - 3' that exclude GGT/GGC/GGG/GAT before the ACC are theoretically thermodynamically stable when forming a Holliday junction, and are also included in the analysis as a way to increase the number of probe sites. After identifying all 40 bp sequences with the form:
   1. 5' –Left\_19\_BP – AG – Right\_19\_BP- 3'
   2. 5' –Left\_19\_BP – TA – Right\_19\_BP- 3'
   3. 5' –Left\_16\_BP – {exclusion of GGT, GGC, GGG, GAT} – ACC – Right\_18\_BP- 3'

The program sequentially aligns these 40 BP sequences with the blastn program against the RefSeq database and ranks the blast results according to their degree of off-target homology. Thus, the MATLAB program uses the native mRNA sequence for blast searches.

1. After this is complete, the 20 BP left and right sequences are reverse-complemented and conjugated to the 24 different H-probe tail sequences that will compliment one of the 24 variable bridges. The form of the conjugation is:
   1. 5’<Reverse comp. of Right 19 BP> <C or T> <Right VB tail>3’
   2. 5’<Left VB tail> <T or A> <Reverse comp. of Left 19 BP> 3’

Note that H-probe tail sequences from PLISH version 1.0 are also included in this step as a legacy function.

1. The newly conjugated H-probe sequences are analyzed for basic oligonucleotide properties such as Tm, delta-G, hairpin formation, and dimerization.
2. A master list of the above data is saved in the form of a text file titled <gene\_name>\_Master\_Probe\_Sheet.txt. We have found this easier to navigate and search the sheet by opening it with a web browser such as Firefox or Opera. The first part of the sheet contains a FASTA file which can be directly copied and pasted into the online NCBI blast search program as a means of double checking the design process if desired. This FASTA file is ordered by degree of off-target homology, with the least off-target homology being at the top. Scrolling further down the sheet will give more detailed results for each probe site, including its blast report, the list of h-probes for the specific probe site complimenting the 24 variable bridges, and thermodynamic properties of each probe.

**Running the script: Option 2**

1. Option 2 can be selected after running part 1 detailed above. This option will merely change the bitscore cutoff for off target homology. It will then recalculate the ranking of the probe sites based on this new score. The script will ask the user to provide the gene name – it must match the gene name typed in step 5 above exactly, case included. After this, a new bitscore value can be entered. The results will be re-saved in the text file described above.

**Running the script: Option 3**

1. If the user decides to move forward with ordering probes designed in part 1 of the PLISH probe design program, re-running the script and choosing option 3 can be useful for producing an excel spreadsheet which makes uploading the probe names and sequences to a commercial vendor easier.
2. After choosing option 3, the program will ask the user to enter the gene name which was used in option 1. The gene name must match the original name exactly, case included.
3. The script will ask you to enter the probe site ID number (this is the number which identifies the start site in the gene) as well as the number of each variable bridge desired separated by spaces. The script will save an excel file to the probe directory of the form: <gene\_name>\_HProbeSheet\_<probe ID numbers>.xlsx.