**Using Environmental DNA (eDNA) to assess biodiversity in NYC waterways**

Billion Oyster Project

In this exercise, you will investigate the fish (and other vertebrate) biodiversity of an aquatic ecosystem or sampling site in NYC using environmental DNA (eDNA) data.

Each of the datasets represents a subset of sequences obtained from water samples from an aquatic site in NYC (including outer boroughs).

1. Select a dataset from Site #1-Site #5 (if you finish your selected site early, feel free to explore another).

2. Go to the following url: <https://blast.ncbi.nlm.nih.gov/Blast.cgi>

Select Nucleotide Blast.

3. Cut and paste one of your sequences in the box labeled “**Enter accession number(s), gi(s), or FASTA sequence(s)”.**

3. Under “Choose Search Set” select “Others” for Database and leave the rest of the settings in the Default value.

This is the option for searching for nucleotide sequences with a nucleotide sequence, but other options (such as searching for translated sequences, searching within the human genome, or searching for really close matches quickly) are available.

5. Under “Program selection” select “Highly similar sequences (megablast)”.

6. Click BLAST button and wait for results (usually several seconds to one minute)

Once the search is done, you can check out which sequences were found that

generated significant alignments with your query sequence by scrolling down the

page. You can also see the alignments with these sequences that the BLAST

algorithm generated as well. There is a graphical representation (near the top of

the results page) that shows where the various hits could be aligned with the query

sequence and how good that alignment is.

The BLAST algorithm calculates similarity scores for local alignments (i.e., the most similar regions between 2 sequences) between the query sequence and subject sequences using specific scoring matrices, and returns a table of the best matches (“hits”) from the database. The hit table includes several useful pieces of information, including the similarity score, query coverage (percent of the query sequence that overlaps the subject sequence), E-value (see below), and max identity (percent similarity between the query and subject sequences over the length of the coverage area). The Graphic Summary displays all the results with how well that matched to the query sequence (your input), the length as well as the color indicate where and how the sequences align.

**What is an E-value?**

From the NCBI website: “The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially as the Score (S) of the match increases. Essentially, the E value describes the random background noise. For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance.

The lower the E-value, or the closer it is to zero, the more "significant" the match is. However, keep in mind that virtually identical short alignments have relatively high E values. This is because the calculation of the E value takes into account the length of the query sequence. These high E values make sense because shorter sequences have a higher probability of occurring in the database purely by chance. For more details please see the calculations in the [BLAST Course](https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html).

The Expect value can also be used as a convenient way to create a [significance threshold](https://blast.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_TYPE=BlastHelp#expect) for reporting results. You can change the Expect value threshold on most BLAST search pages. When the Expect value is increased from the default value of 10, a larger list with more low-scoring hits can be reported.”

How many hits did you get?

For each sequence results, scroll to the table of GenBank matches. Read through the top 10-20 matches.

a. What is the top identity of the species that is best matched with the unknown sequence? (Record the accession number – the last column listed in the table).

Look in particular at the E-value and the Ident columns.

b. Are there multiple species with the same scores for the E-value and Ident columns? What does this tell you?

c. For each record, click on the Accession Number to pull up the record for that sequence in Genbank. This record will tell you the organism that it was sequenced from, the publication or project it is associated with, and sometimes other relevant info like collection locale. Click on “Taxonomy Reports” if you want to learn more about the organism that matched the query sequence. Now, repeat the BLAST exercise for the eDNA sequences from your site in NYC!

d. Google the top hits for species to find out more about their habitats and environmental requirements.

e. Repeat for all the sequences from your site. Do you notice a difference in the number of exact matches in shorter versus longer sequences?

f. Once you have a list of potential species matches, make an educated guess about the location or habitat type that this water sample was taken from.