**Picking the best match for your query sequence**

1. **What is the primary citation for the BLAST program?**

See under the BLAST results:



Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997), "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", Nucleic Acids Res. 25:3389-3402.

1. **How many sequences are there in the database that was used for your search?**

Click on the ‘Search summary’ link:



A new box with information on the database will open. The information below is from December 7th, 2018:



49,275,898 as of 12/7/18; note this number will likely change as the databases are growing larger all the time

1. **What does the color code on the graphical interface represent?**



Color key represents alignment scores – higher score = higher similarity. Red means it is the most similar, followed by pink.

1. **What is the bit score, E-value and query coverage values of the first ten hits of the BLAST results? Copy the table that appears to capture the information.**

****

The bit score or Max score is calculated from the alignment and normalized using the scoring matrix used for the BLAST search (https://www.ncbi.nlm.nih.gov/books/NBK62051/)

1. **Among these ten results, which represents an ortholog in another close species?**

*Bacteroides salanitronis* DSM 18170 (highlighted in table in Q4 – other hits above this are the same species)

1. **There are some entries with significant E-values (low E-values). Are these all best matches to the entire query sequence? How would you pick the best match between the query and subject?**

The first three hits in the table match genes in different strains of *Bacteroides thetaiotamicron*. The fourth hit matches genes in a different species. All these hits have high matching scores, high query coverage and low E-values. These criteria together are used to pick the best matches.

The fifth hit titled *Bacteroides thetaiotomicron neopullulanase* is not a good match due to a low query coverage (20%).

1. **How would you identify a closely related species to your query sequence?**

The criteria used for identify closely related species is High bit score, low E-value and a high query coverage. In addition, query and subject length should match with few gaps in the sequence alignments.

