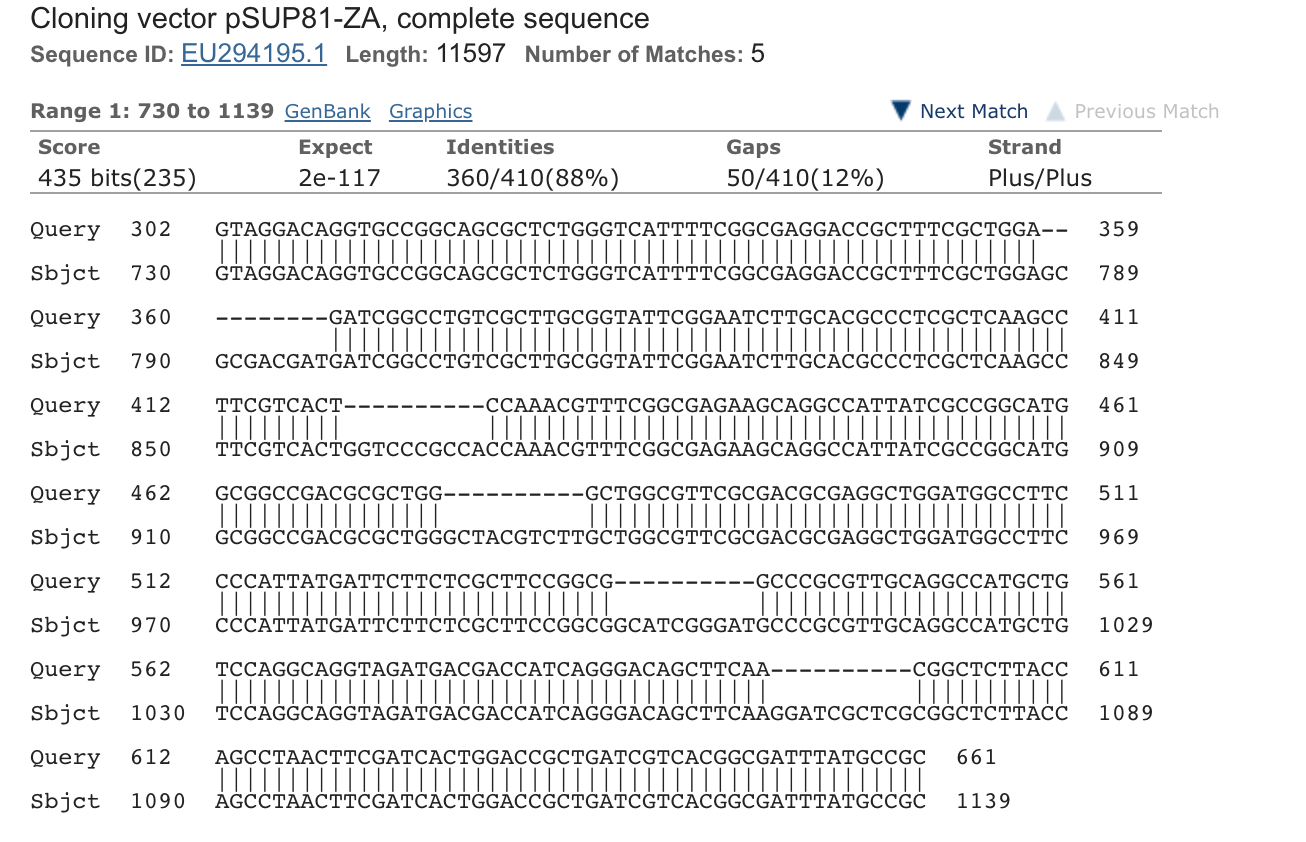
Lab 3

2.1) BLAST uses an expected value (E-value) and a bit score for each alignment pair. These two scores give users a sense of whether there is an underlying biological relationship. The bit score reveal the how strong the alignment between sequences are. The higher the bit score, the better the alignment. Bit scores are calculated by using a substitution matrix which takes into consideration gaps and assigns a score based on similarity between two sequences. The E-value illustrate how significant the pairwise alignment is. The lower the E-value, the more significant the hit. The E-value is calculated using the formula E=Kmne-λS where m and n are the sequence lengths, K is space size and lamba scoring system, and S is the score and the P-value is calculated by using P = 1 – e-E.

2.2) BLAST is a tool designed for two nucleotide or two protein sequences to search for homologous sequences. BLAST is used to search for sequences from the same species so using database search is unnecessarily time consuming. BLAST 2 takes advantage of BLAST algorithm for sequence alignment and utilizes substitution, insertions, and deletions. BLAST 2 compares sequences of different strains and finds multiple local alignments. The difference between BLAST and PSI-BLAST is the scoring system. PSI-BLAST captures patterns of alignment and stores it in specific positions so it is largely dependent of position of alignment where as BLAST is independent of positions when calculation alignment scores. BLAST should be should for quick search because of the rapid speed and with sequences that differ only in a few nucleotides. It does not provide efficient accuracy. BLAST 2 is used to find multiple local alignment between sequences and is useful in studying genome variations. However, it is not useful to search for motifs and aligning metabase sequences.

2.3) The E-value for the given query is 2e-117. This low value indicates that the hit is very significant. It means that the similarity has a 2 in 10117 chance of occurring by chance alone and may exhibit biological significance.



For my calculation I used:

E = mn2-S  =

1201\*11597\*2^-435

0

Assuming that n is very large, the expected value is approximately 0 which is comparable to BLAST’s E=value of 2e-117 which is negligible.

3) A HSP gets reported multiple times because when comparing a query string to a database, a kmer from the query can appear in multiple positions of a database string so when the query is found and alignment matching begins, the observed kmer can appear once in a query but can appear twice or more in the database. So now matching can be done with the query string and 2 different positions of the database string. Hence, even if using the first location of k-mers, HSP can get reported multiple times if it appears multiple times in the database.

Longest Lengths

|  |  |  |
| --- | --- | --- |
|  | K = 3 | K = 4 |
| L = 6 | 6 | 6 |
| L = 7 | 7 | 7 |