An Introduction to NCBI BLAST — Worksheet

Wilson Leung

This worksheet contains questions related to the walkthrough from "An Introduction to NCBI BLAST".

Question 1. Which of the hits under "Description" is the best hit for your unknown sequence (provide its name and accession number)? Use the metrics provided in the search results to support your reasoning. (*Note: We will ignore all computationally predicted genes and focus only on experimentally verified genes. This is because the experimentally verified genes are the only ones that we can be sure actually exist.*)

Question 2. Fill in Table 1 of the worksheet with information for each alignment block from the *blastn* search. Make sure to include the total number of mRNA base pairs in the table title. *(Note: If you do not have enough ranges to fill all rows, leave the remaining rows empty. Make sure your ranges are sorted by subject position.)*

Table 1. *blastn* results for the unknown sequence and the *D. melanogaster legless* mRNA
Total number of mRNA base pairs: _____bp

Subject Position Range	Query Position Range	E-value	% Identity

Question 3. Use data from Table 1 to assert whether the entire *D. melanogaster legless* mRNA aligns to our unknown sequence. That is, provide evidence to support or reject whether the sequences matching *D. melanogaster legless* mRNA in your unknown sequence show collinearity.

Question 4. Write a hypothesis indicating what gene from what species is homologous to a region of your *D. yakuba* unknown sequence.

Question 5. How do RefSeq records with the "XM_" prefix differ from records with the "NM_" prefix? How do RefSeq records with the "NM_" prefix differ from records with the "NP_" prefix?

Question 6. Fill in Table 2 of the worksheet with information for each alignment block from the *blastx* search. Make sure to include the total number of protein amino acids in the table title. *(Note: if you do not have enough ranges to fill all rows, leave the remaining rows empty. Make sure your ranges are sorted by subject position.)*

Table 2. *blastx* results for the unknown sequence and the *D. melanogaster* legless protein

Total number of protein amino acids: _____aa

Subject Position Range	Query Position Range	E-value	% Identity	% Positives	Frame

Question 7. Use data from Table 2 of your worksheet to assert whether the entire *D. melanogaster* legless protein aligns to your unknown sequence. First add what support you find, using similar reasoning as in Question 3. Then note any discrepancies, clearly distinguishing between supporting and conflicting evidence.

Question 8. Perform *tblastn* searches of each CDS of *legless* (query) against the unknown sequence (subject). Fill in Table 3 on your worksheet based on the *tblastn* search results.

Table 3. *tblastn* results for the coding sequences (CDS) for *legless* in *D. melanogaster* and the unknown sequence

CDS #	FlyBase ID	Query Position Range	Subject Position Range	E-value	% Identity	% Positives	Frame

Question 9. Use data from Table 3 to explain whether all CDS of the *D. melanogaster legless* gene are accounted for in your unknown sequence. Explain how the CDS *tblastn* alignments resolve the discrepancies discussed in Question 7.

Additional exercise

Question 10. Based on the *blastn* search result, where is the best match for lgs:1 in the unknown sequence? What are the E-value and the percent identity for this *blastn* alignment? Based on the *tblastn* result for CDS 1_9469_0 and the *blastn* alignment for lgs:1, where is the 5' UTR for *lgs* in the unknown sequence?