BLASTing through the kingdom of life

Information for teachers

Description:

In this activity, students copy "unknown" DNA sequences and use them to search GenBank, the main database of nucleotide sequences at the National Center for Biotechnology Information (NCBI). All of these sequences originally came from GenBank so each sequence will have at least one match. These sequences came from a wide variety of organisms including viruses, bacteria, plants, mammals, frogs, humans, and other creatures. All these sequences code either for a protein or an RNA molecule and are missing introns, making the results easier to interpret.

Helpful hints:

1. The question that confuses students the most is whether these sequences are expressed. The answer is yes for all the sequences. Gene expression is defined as the process of transcribing a gene to make a molecule of RNA. Since all these sequences come from either mRNA or rRNA, all of these sequences are expressed.

2. It's best to have the students look at the tutorial first then do the blast search on their own. They can do this at home or anywhere with internet access, since the tutorial is freely available on line.

3. To access an updated answer key, go to DigitalWorldBiology.com and sign up for an Educator account.

Materials:

- An Internet connection.
- Make a bookmark in your web browser to the NCBI web site: https://www.ncbi.nih.nlm.gov
- Make a bookmark to the **BLAST for beginners** tutorial at Digital World Biology: http://digitalworldbiology.com/tutorial/blast-for-beginners
- The **BLAST for beginners** tutorial has links to an interactive tutorial that shows how to do a blast search and interpret the results and a set of 16 "unknown" sequences that students can identify along with one example that shows how the questions might be answered.

Gotchas

- The NCBI web site changes frequently. The BLAST tutorial pages may look somewhat different than the pages at the NCBI.
- Some of the questions below are only appropriate for some types of sequences. For example, bacteria are single-celled organisms that do not have tissues. If your gene is a bacterial gene, it is unlikely that you will find tissue-specific expression.
- This is real research so some questions may not have answers.

Information for students

Instructions:

In short, you will copy one sequence from the data set, use blastn to identify it, and use the information from your search to answer the questions below. Instructions for copying and pasting sequences are provided with the data set. Instructions for using BLAST are in the BLAST for beginners tutorial.

Materials:

- An Internet connection.
- Make a bookmark in your web browser to the NCBI web site: https://www.ncbi.nlm.nih.gov
- Make a bookmark to the **BLAST for beginners** tutorial at Digital World Biology: http://digitalworldbiology.com/tutorial/blast-for-beginners
- The **BLAST for beginners** tutorial has links to an interactive tutorial that shows how to do a blast search and interpret the results and a set of 16 "unknown" sequences that students can identify along with one example that shows how the questions might be answered.

Gotchas

- The NCBI web site and the contents of NCBI databases change frequently. The BLAST tutorial pages may look somewhat different than the pages at the NCBI.
- Some of the questions below are only appropriate for some types of sequences. For example, bacteria are single-celled organisms that do not have tissues. If your gene is a bacterial gene, it is unlikely that you will find tissue-specific expression.
- This is real research so some questions may not have answers.

Questions: Be sure to include the source of the information along with your answer. In this case, the source will be the database or web page that provided the information.

1. How long is the sequence that was used to search the database?
   Hint: This sequence is called the "query" sequence because you used it to query the database. (The phrase "to query" means to "ask a question." When you do a blast search you are asking the database a question about its contents.)

2. Which sequence matches your query the best? What data support this conclusion?
   Hint: Refer to the slide in the BLAST tutorial that discusses the E value.

3. What organism was the most likely source of the sequence?
   Hint: Refer to the BLAST tutorial to find an overview of the GenBank nucleotide record. If sequences from more than one organism match your query, use the E value, the % identity and fraction of the query matched to determine the best match. If all the nucleotides in a sequence match, the % identity will be 100%. If the query sequence is as long as the database sequence, then 100% of the query will be aligned to the subject sequence. In some cases, you may find multiple sequences are identical to the query and match the same extent. If multiple sequences are identical, then the correct answer will be that multiple sequences are identical to the query. You can also use the total blast score.

4. What is the common name for this organism?
   Hint: Refer to the GenBank nucleotide record. It may also help to look at the Taxonomy database. The BLAST tutorial shows where to find this link.

5. What phylum contains this organism?
   Hint: Refer to the taxonomy database. The BLAST tutorial shows how to find the link. If your sequence is a virus, this question doesn't apply exactly, but you can still describe the kind of virus and the kind of genome it contains.

6. What is the accession number for the best-matching sequence?
   Hint: Refer to the summary table in your BLAST results. The BLAST tutorial shows where to find this.

7. Estimate the number of sequences in the table with an E value less than 0.001.
   Hint: Refer to the blast results. Note – by default, the number of sequences in the table is limited to 100. If all your sequences show an E value of 0.0, you don't have to count them.

8. If possible, give the names of three different organisms that have sequences with significant E values. Pick the three where the values are the most significant.
   Hint: Refer to the BLAST tutorial slide on E values for a description.

9. Look at the first matching sequence in the table. For that subject sequence, determine the length of the alignment, in nucleotides, and the fraction of nucleotides that match your query sequence.
   Hint: Refer to the BLAST tutorial slide on interpreting values in the results table.

10. Draw a picture to represent the best alignment between the two sequences, with the query sequence on top, and include the starting and ending map positions for both sequences.
    Hint: Your picture should look similar to the alignment between the query and subject sequence in the graph of the blast hits.

11. Use the information and links from the GenBank nucleotide record to see if you can find a function for the protein or RNA specified by your DNA sequence.

Hint: Read the entire title of the sequence record to see if you can find the name of a protein or type of RNA. You can also look in the GenBank nucleotide record for the word "product." You can often use the name of protein or RNA to look up its function.

12. Is this sequence expressed? How do you know?
    Hint: Gene expression includes the processes of transcription (making RNA) and translation (making a protein).

    Do you see either of these molecules mentioned in the database record? Find the "mol_type" field. What kind of molecule was sequenced? Look for the "product" field. What kind of product does this sequence make? See the BLAST tutorial page on the GenBank record for help locating this information.

    Scan the nucleotide record for links or references to proteins or RNA molecules. If these exist, then this gene is likely to be expressed.

    GenBank nucleotide records also tend to include amino acid sequences for any proteins that might be produced when a gene is expressed. These sequences can be found in the CDS (coding sequence) section of the nucleotide record.

    Watch out for genome sequences or sequences of larger molecules. If your sequence matches a longer molecule of DNA, then, you might see references to multiple proteins. If that's the case, look in the alignment portion of the BLAST results to see where your sequence matches, then focus on the matching region in the nucleotide record.

13. If your sequence is expressed, where is this gene expressed?
    Hint: In multicellular organisms, different genes can be expressed in different tissues.

    First, determine if your sequence is from a multicellular organism. If your sequence is from a virus, yeast, or bacteria, this question doesn't apply.

    If your sequence is from a multicellular organism, look in the nucleotide record to find the tissue type and possibly the cell type.

14. Is there a specific time during development when this gene is expressed?
    Hint: If your sequence is from a multicellular organism, the nucleotide record may include a reference to the developmental stage of the sample material. This would be abbreviated as "dev_stage." The tissue type or cell type might also tell you when this gene expressed.

15. Is anything known about factors that cause your sequence to be expressed?
    Hint: The title of the submission is a good place to start. PubMed records and the Entrez Gene database may also be helpful.

Answers for the example sequence

1. How long is the sequence that was used to search the database?

   840 nucleotides
   Information source: BLAST format page

2. Which sequence matches your query the best? What data supports this conclusion?

   The mostly likely identity for this sequence is the human tyrosinase gene.
   The E value is close to zero, suggesting a low probability that the match would occur by chance.
   Information source: BLAST results.

3. What organism is the mostly likely source of the sequence?

   Homo sapiens (common name = human)
   Information source: The description of the sequence in the table of blast results. This could also be found in the GenBank nucleotide record

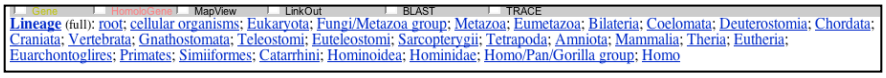
4. What is the common name for this organism?

   Human
   Information source: The GenBank nucleotide record

5. What phylum contains this organism?

   Chordata.

   Information source: The Taxonomy database. I got this information by finding the scientific name in the ORGANISM field of the GenBank nucleotide record. Clicking “Homo sapiens” took me to the taxonomy database. Then I held the pointer over each of the taxonomy levels (shown below). A tag appears that identifies the level (kingdom, phylum, order, etc.)



6. What is the accession number for the best-matching sequence?
   NM_000372
   Information source: The blast results table and the GenBank nucleotide record.

7. Estimate the number of sequences with an E value less than 0.01.

   Over 100

   Information source: the blastn results table – the settings on the blastn program limit the number of descriptions to 100. All the descriptions had a significant E value.

8. If possible, give the names of at least three different organisms with significant E values. Record the name of the organism, the common name, and the E value.

    - Homo sapiens, E value close to 0, total blast score 1748, 95% identical in aligned region
    - Pan paniscus, (Bonobo chimp) E value close to 0, total blast score 1730, 94% identical in best aligned region
    - Pan troglodytes, (chimpanzee) E value close to 0, total blast score 1709, 94% identical in best aligned region

    Information source: the blastn results table

9. Look at the first matching sequence in the table. For that subject sequence, determine the length of the alignment, in nucleotides, and the fraction of nucleotides that match your query sequence.

    The aligned region is 1104 nucleotides long and 95% of the nucleotides match the query sequence.

10. Look at the alignment to the first matching sequence and determine the length of the alignment and the fraction of nucleotides that match your sequence. Draw a picture to represent the alignment between the two sequences and include the starting and ending map positions for the both sequences.

    Query 1 ---------- 840
    Subject 961 ----------2064

11. Use GenBank, PubMed, Gene, and UniGene records to find the possible function of the protein that's specified by your DNA sequence. Describe what's known about the role of this protein in the organism that provided the DNA.

    This protein is involved in making the pigment found in skin. Mutations in this protein are associated with albinism.

    Information source: The Gene database was the source of this information. I also used the links on the right hand side of the nucleotide record.

12. Is this sequence expressed? How do you know?

    Yes, this sequence is expressed. The description says that this is mRNA. The nucleotide record also says the molecule that was sequenced was mRNA. (*Students might be expected to explain why the presence of mRNA shows that a gene is expressed*).

    Information source: The blast results table and the GenBank nucleotide record.

13. If your sequence is expressed, where is it expressed?

    The titles of the first two articles in the GenBank nucleotide record mention melanoma cells, those cells are present in skin cancer. Other titles talk about pigmentation and skin cancer. That leads me to conclude that this gene is expressed in the skin.

    Many of the titles also contain the word "oculocutaneous." Looking up the definition of oculocutaneous shows that the sequence has something to do with eyes, so it's likely that the sequence is also expressed in eyes.

    To investigate expression further, sometimes it's possible to find information in the Unigene database. I searched all the databases with word "tyrosinase" and clicked the link to UniGene database.

    I picked the human Unigene record and found information about where this gene is expressed. The results for the human tyrosinase gene are shown in the table on the right.

    Tissues are listed are in the first column, the number of transcripts (RNA molecules) per million is shown in the next column, a dark dot is shown in the third column to indicate the relative amounts of expression, and the last column shows the number of transcripts that match out of the total number tested.

Breakdown by Tissue

	Hs.503555	
bladder	0	0 / 21352
blood	0	0 / 77360
bone	0	0 / 54806
bone marrow	27	1 / 36016
brain	0	0 / 462807
cervix	0	0 / 40857
colon	0	0 / 177778
eye	41	7 / 167922
heart	0	0 / 57999
kidney	14	2 / 137517
larynx	0	0 / 27036
liver	0	0 / 130428
lung	0	0 / 286629
lymph node	0	0 / 127387
mammary gland	43	6 / 138153
muscle	9	1 / 108250
ovary	0	0 / 94739
pancreas	0	0 / 196747
peripheral ...	0	0 / 24783
placenta	0	0 / 233561
prostate	0	0 / 132437
skin	310	51 / 164361
small intes...	0	0 / 14023
soft tissue	0	0 / 23646
spleen	0	0 / 19096
stomach	0	0 / 107451
tongue	0	0 / 28525
testis	0	0 / 135901
thymus	0	0 / 6782
uterus	0	0 / 179761
vascular	0	0 / 25627

    These data show that this gene is expressed mostly in the skin, but also in the bone marrow, eye, kidney, mammary gland, and muscle.

    Information sources: the GenBank nucleotide record, UniGene

14. Is there a specific time during development when this gene is expressed?

    The results from the Expression Profile in UniGene show that it's expressed mostly in the embryo and a little in adults.

    Information sources: UniGene, Entrez Gene

Breakdown by Developmental Stage

	Hs.503555	
embryo	14	8 / 549513
juvenile	0	0 / 57387
adult	5	5 / 962706

15. Is anything known about factors that cause your sequence to be expressed?

    I searched the Gene database for human tyrosinase and found that increased expression was stimulated by exposure to the sun.

    Information sources: Gene database, Blast results