

BLASTing through the kingdom of life

www.geospiza.com/education

Sandra Porter, Ph.D.
Geospiza, Inc.
sandy@geospiza.com

Materials:

- An Internet connection
- A web browser. I recommend Firefox (www.mozilla.org).
- Data set and instructions. This data set contains 16 unknown sequences and one example.
 1. Go to <http://www.geospiza.com/education/materials.html>
 2. Look under Data Sets for BLAST for beginners
 3. Select taxonomically diverse sequences

Brief instructions:

In short, you will copy one of the sequences 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.

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.

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 ask a question (or query) of the database.
2. What is the most likely identity of this sequence? What data supports 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 more than one organism matches, look at the E values to determine the most likely match.
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.

BLASTing through the kingdom of life

www.geospiza.com/education

5. What phylum contains this organism?
Hint: Refer to the taxonomy database. The BLAST tutorial shows how to find the link.
6. What is the accession number for the best-matching sequence?
7. Is this sequence expressed? How do you know?
Hint: Gene expression includes the processes of transcription (making RNA) and translation (making a protein). Determine if either of these molecules is described in the sequence record.
8. If your sequence is expressed, where is this gene expressed?
Hint: The sequence record usually indicates the source of the material that was sequenced. In some cases, however, you will probably need to look at other sources of information. Some places to look are the title of the submission, PubMed records (you might need to look at more than one record), the UniGene database, and the Gene database. If your BLAST results show boxes with U's or G's (shown below) those are links to the UniGene and Gene databases respectively. The Expression Profile in the UniGene database is linked to a set of tables that show the tissues where mRNAs have been found and the developmental stage where they were found.

io sapiens tyrosinase (oculocu... [1665](#) 0.0 

9. Is there a specific time during development when this gene is expressed?
Hint: See the hint for question 8.
10. 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 are also helpful; see the hint for question 8.
11. Estimate the number of sequences with an E value less than 0.01.
Hint: Refer to the blast results.
12. If possible, give the names of three different organisms with significant E values. If organism is represented, then write down the name of that organism.
Hint: Refer to the BLAST tutorial slide on E values for a description.
13. Look at the first matching sequence; 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.
14. 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.

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. What is the most likely identity of this sequence? 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 GenBank nucleotide record

4. What is the common name for this organism?

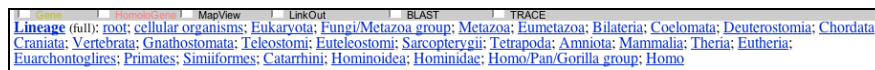
Human

Information source: The GenBank nucleotide record

5. What phylum contains this organism?

Chordata.

Information source: The Taxonomy database. I got this information by clicking the link from the GenBank nucleotide record to the taxonomy database. Then I held the pointer over each of the taxonomy levels (shown below). A yellow tag appears that identifies the level (kingdom, order, etc.)



6. What is the accession number for the best-matching sequence?

NM_000372

Information source: The GenBank nucleotide record.

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

Yes, this sequence is expressed. I know this because 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 GenBank nucleotide record.

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

The titles of articles in the GenBank record show that this sequence is expressed in the skin (human skin melanocytes) and melanoma producing cells. Looking up the

BLASTing through the kingdom of life

www.geospiza.com/education

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, I clicked the U next to the sequence to obtain the record from the UniGene database.

Clicked the Expression Profile (in the middle of the UniGene record) leads to information about where this gene is expressed. The results for the human tyrosinase gene are shown in the table on the right.

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

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.

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

9. 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

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

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

I clicked the G to look at the Entrez Gene database. In this record, I found that increased expression was stimulated by exposure to the sun.

Information sources: Entrez Gene, Blast results

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

Over 100.

BLASTing through the kingdom of life

www.geospiza.com/education

Information source: the blastn formatting page – the settings on the blastn program limited the number of descriptions returned to 100 and all the descriptions given had a significant E value.

12. If possible, give the names of at least five 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
Gorilla gorilla, (Gorilla) E value close to 0
Pan troglodytes, (chimpanzee) E value close to 0
Pan paniscus, (Bonobo chimp) E value close to 0
Pongo pygmaeus (orangutan), E value close to 0

Information source: the blastn results

13. Look at the first matching sequence; 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 1 ----- 840

14. 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.

(Students would be expected to write at least a paragraph describing some of the information that's known about the gene).

Information source: Entrez Gene was the source of this information.