Introduction to BLAST using human leptin

Developed by Justin R. DiAngelo (Penn State Berks) and Alexis Nagengast (Widener University)
Revised by Wilson Leung (Washington University in St. Louis)

What is BLAST?[1]

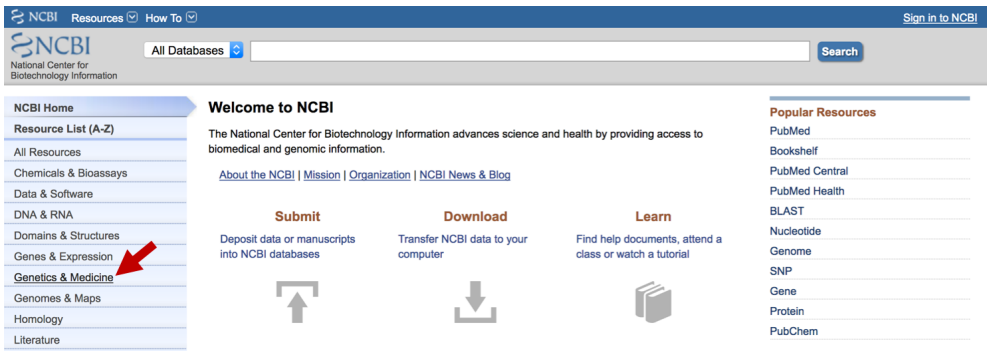
BLAST stands for **B**asic **L**ocal **A**lignment **S**earch **T**ool and it is a program that reports regions of similarity (at the nucleotide or protein level) between a query (your input) sequence and sequences within a database. BLAST uses a robust statistical framework that determines if the alignment between two sequences is statistically significant (i.e. has a low probability of the reported alignment being produced by chance alone). The ability to detect sequence similarity allows scientists to determine if a gene or a protein is related to other known genes or proteins in the same species or between species.

The theory of evolution is based on all organisms descending from common ancestors by speciation. At the molecular level, an ancestral DNA sequence diverges over time (through accumulation of point mutations, duplications, deletions, transpositions, recombination events, etc.) to produce diverse sequences in the genomes of living organisms. Such sequences are classified as *homologs* if they come from the same ancestral gene.

Mutations in genes with an important biological function have a higher probability of being harmful to the organism and are less likely to become fixed in a population. Such sequences are said to be under *negative* selection, which causes them to be *conserved* against change over time. Therefore, it is expected that two homologous copies of a functional sequence will show a higher degree of sequence conservation (observed as base-by-base similarity at the nucleotide level) than either two unrelated sequences or two sequences that are not under strong negative selection. This similarity is the "signal" detected by a BLAST search.

Obtaining sequence using NCBI[2]

The National Center for Biotechnology Information (NCBI) is a public database that houses molecular biology information including sequences from thousands of different species from mammals to fungi. We will explore some of the basic functionalities of the NCBI web site using leptin (*LEP*) — a gene that has been found to contain mutations associated with severe obesity and the development of type 2 diabetes. First, open a web browser and navigate to the NCBI web site at https://www.ncbi.nlm.nih.gov. To get information on obesity, click on the Genetics and Medicine link on the left, scroll about 1/3 of the way down the page and click on the Genes and Disease link. Scroll down the page and click on the "Nutritional and Metabolic Diseases" link, then click on the "Obesity" link to find a non-technical description of the hormone leptin and its role in weight control. On the right panel, you will find links to other parts of NCBI that contain more information



about this disease. For example, you can access the corresponding gene record in the OMIM database (a catalog of human genes and disorders) through the "OMIM" link.

1. Based on the information on this page, how does *leptin* control feeding?

To obtain the sequence for the human *LEP* gene, go back to the NCBI homepage at https://www.ncbi.nlm.nih.gov/. At the top of the page, use the pull-down menu next to "Search" and select "Gene". Enter "LEP homo sapiens" into the text box and click Search.

Note that this search menu is also available at the top of the NCBI Bookshelf page. Hence you could scroll to the top of the page, click on the drop-down box to select "Gene" and then search for "LEP homo sapiens" directly. You can also access the Entrez gene record through the "Entrez Gene" link under the "Gene sequence" section on the right panel.

This search produces 51 results, with the first entry being the gene record for human *leptin*. The remaining results may mention *leptin* in their detailed summary. Even before we click on the *leptin* entry, we can already obtain some useful information about the *leptin* gene from the search results page. For example, the chromosomal location and OMIM entry number for *leptin* are shown.

2. According to these search results, on which chromosome is the *leptin* gene located?

Click on the first match to the human *LEP* gene to learn more about this gene and its sequence. The Entrez Gene record for *LEP* shows lots of detailed information about gene structure and function.

The first section of the Entrez Gene record is the Summary section and it contains:

- The official symbol and name approved by HUGO Gene Nomenclature Committee (HGNC)
- Other synonyms that have been used to describe this gene
- The organism the gene is from and the lineage of that organism
- Links to external databases (*e.g.*, HGNC and Ensembl) that contain similar information
- Link to the NCBI's Online Mendelian Inheritance in Man (MIM) that provides comprehensive information aimed for the medical and scientific research community
- RefSeq status. The NCBI Reference Sequence Database (RefSeq) is a comprehensive, curated database of non-redundant sequence records. The accession number for a RefSeq record begins with two characters, follow by an underscore. This prefix denotes the type of sequence record:
    - chromosomes: NC_
    - genomic regions: NG_
    - mRNA: NM_
    - proteins: NP_

The "Genomic regions, transcripts, and products" section has a map of the gene with links to the sequence. The sequence record can be retrieved in many different formats, including FASTA (a simple text format which contains only the sequence — easiest for BLAST searching) or GenBank (more descriptive version with the sequence listed at the end). The *leptin* gene is on the positive strand of the chromosome (meaning that if the chromosome was oriented from left to right, your gene would be on the strand running from 5' to 3'). Specifically, the *leptin* gene can be found at 128,241,201-128,257,629 of human chromosome 7 in the current assembly (GRCh38.p7).

3. How do you know that the *leptin* gene is on the + strand of the chromosome it's located on?

4. The picture depicts two annotated transcripts (XM_005250340.4 and NM_000230.2). What is the difference between the RefSeq record that begins with the XM_ prefix and the record that begins with the NM_ prefix? (Hint: Examine the "COMMENT" section of the GenBank record or the RefSeq FAQ page.) Both mRNA records have 4 green boxes, 2 light green in color (1 on each end) and 2 dark green. What do you think each of those boxes refers to? (Hint: click on the green feature; this will give you more information on the *LEP* gene.)

Sequences are most conserved between species at the amino acid level and this is what we will use in our BLAST searches. The protein sequences for the two annotated transcripts are available under the "NCBI Reference Sequences" section. Note that there are multiple options to obtain the protein sequence for the leptin precursor, including the RefSeq protein database (NP_000221.1), the Consensus CDS (CCDS) database (CCDS5800.1), the UniProtKB/Swiss-Prot database (P41159) and the UniProtKB/TrEMBL database (A4D0Y8).

To do your own BLAST search, scroll down to the "NCBI Reference Sequences" section and click on the "NP_000221.1" link.

5. Why do you think you want the protein sequence as opposed to the nucleotide sequence?

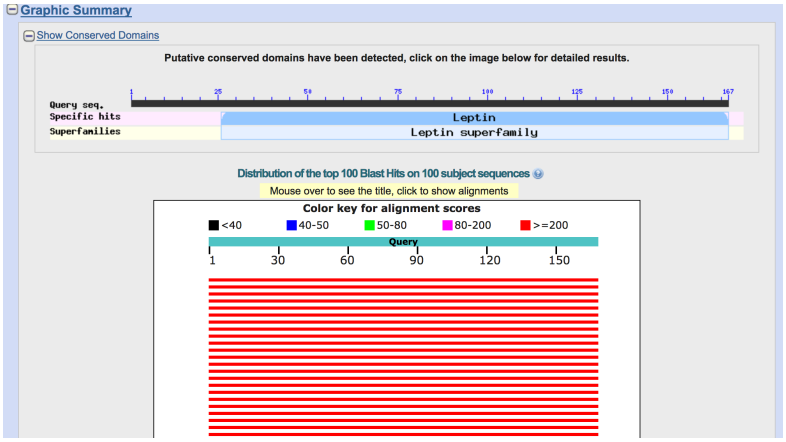
This screen gives you information on the LEP protein sequence. Click on the FASTA link directly beneath the gene name at the top of the screen. This gives you the FASTA definition line (as indicated by >) followed by the one letter amino acid sequence. The definition line varies but it always begins with the accession number of the sequence record (*e.g.*, NP_000221.1), followed by the gene name and the organism. Copy the sequence including the definition line for use in a BLAST search.

Performing a BLAST search[2]

Open a new tab in your web browser, go back to https://www.ncbi.nlm.nih.gov/ and click on the "BLAST" link in the "Popular Resources" section on the right. There are many different options for a BLAST search and the option you choose depends on the sequence you have and the degree of conservation you are trying to detect. Start by clicking on the "Protein BLAST" image under the "Web BLAST" section and paste your sequence into the "Enter Query Sequence" text box. Alternatively, you could enter the accession number (i.e., NP_000221.1) into the text box. If you think you will be doing many searches at once, you can enter a Job Title to distinguish one from the other; if not, the FASTA definition line will be used. Among the databases available, the non-redundant protein database is the most comprehensive; it contains sequence records from most of the other databases listed in the drop-down box:

- **Non-redundant protein sequences (nr):** non-redundant merge of protein sequences from the NCBI's GenBank coding sequence translations, RefSeq proteins, Swiss-Prot, PDB, PIR (Protein Information Resource at Georgetown University Medical Center), and PRF (Protein Research Foundation)
- **Reference proteins (refseq_protein):** manually curated database from NCBI
- **Model Organisms (landmark)**: protein sequences from 27 genomes that cover a wide taxonomic range (from bacteria to human)
- **UniProtKB/Swiss-Prot (swissprot):** manually annotated and reviewed section of the UniProtKB database
- **Patented protein sequences (pat):** proteins from the Patent division of GenBank.
- **Protein Databank proteins (pdb):** amino acid sequences derived from 3-dimensional structures from the Brookhaven Protein Data Bank
- **Metagenomic proteins (env_nr):** protein sequences from environmental samples
- **Transcriptome Shotgun Assembly proteins (tsa_nr)**: transcripts assembled from ESTs and second generation sequencing reads (*e.g.*, Illumina)

Under the "Choose Search Set" section, select the "non-redundant protein sequences (nr)" option from the "Database" drop-down menu. Then click on the blue BLAST button to start your search. You will be taken to a page that will update every few seconds until the BLAST search is complete.



Before performing the blastp search, NCBI BLAST will search the query sequence against the Conserved Domains Database (CDD). If the protein sequence contains a conserved domain, a graphical representation of conserved protein domains will appear. You can click on the conserved domain (*e.g.*, Leptin superfamily) to learn more about the conserved domain.

When your search is complete, you will get a graphical output of the database entries (subjects) that align with your input leptin sequence (query) on a color scale. Features with highly statistically significant alignments will appear in red in the graphical output while features with the lowest statistical significance will appear in black. Move your mouse so that it is above the first red bar below the scale. A tooltip will appear which shows the sequence title. Click on the feature to see a more detailed description of the match (*e.g.*, score, E-value, accession), as well as a link to navigate directly to the corresponding alignment.

Scroll down to the "Descriptions" section. The first entry in the "Descriptions" table shows the best match to the query sequence (leptin precursor [*Homo sapiens*]). The match has a score (S) of 340 and an expect (E) value of 2e-118. Alignments with smaller E-values are more statistically significant and are less likely to have arisen by chance. Specifically, the E-value denotes the number of times we expect to see an alignment with scores equal to or greater than S when we align random sequences against each other. By default, NCBI BLAST will report matches with an E-value of less than 10. The last column of the "Descriptions" table contains the accession number for the match (NP_000221.1), and the link allows you to access the corresponding GenPept record.

The second entry in the "Descriptions" table shows a match to the leptin precursor in chimpanzee (*Pan troglodyte*). This match has a total score of 339 and an E-value of 6e-118. Because the accession number for this match (NP_001180601.1) begins with "NP_", this protein record is from the RefSeq database.

Click on the "leptin precursor [*Pan troglodytes*]" link under the "Description" column to navigate to the alignment between your input leptin sequence (query) and the leptin precursor in chimp (subject). The header in the alignment output shows that the chimp leptin precursor sequence has a total length of 167 amino acids. The alignment has an E-value of 6e-118 and there is a single amino acid substitution between the human and chimp leptin precursors (from V to M). This high degree of sequence similarity is unlikely to be caused by chance alone.

The alignment is shown as three separate lines; the first (Query) line is our human sequence, the third (Sbjct) line is the chimpanzee sequence and the middle (Match) line represents a comparison of the two sequences (if the amino acids are the same between the two sequences, that amino acid will be included in the middle line). The value in the "Identities" field corresponds to the number of identical amino acids between the query and the subject sequences. In this case, this field (166/167) indicates that, of the 167 amino acids in the alignment between the human and chimp sequences, 166 of the amino acids (99%) are identical. The one difference between these two sequences can be found at position 94 of the query (as a + in the Match line), where the human sequence is a valine (V) and the chimp is a methionine (M).

6. What is the relationship between V and M that warranted a + designation?

Click on the "Descriptions" link in the gray toolbar to navigate to the "Descriptions" table. Scroll down to the RefSeq record (with the accession number NP_001277830.1) for the leptin precursor in *Bubalus bubalis* (water buffalo). The E-value is still very highly significant at 3e-96. Click on the link under the "Description" column to navigate to the alignment. The *Bubalus bubalis* leptin precursor protein has the same length (167aa) as the human leptin precursor protein, and the entire protein is included in the alignment. However, there are many more changes and the percent identity is substantially lower than the alignment to the chimp leptin precursor protein.

7. What is the percent identity and percent positives (aka percent similarity) of the *Bubalus bubalis* leptin sequence compared to the human sequence?

You can also perform searches against the database of specific organisms. Go to the BLAST page at https://blast.ncbi.nlm.nih.gov/Blast.cgi and enter the name of the organism in the text box under the "BLAST Genomes" section to search its databases. For example, enter "chimpanzee (taxid:9598)" into the text box and then click on the "Search" button to search the chimp assembly. To search for protein matches, click on the "blastp" tab, paste the human leptin precursor sequence into the "Enter Query Sequence" text box, and then verify that the "RefSeq protein" option is selected under the "Database" field. Because the database is much smaller (with 60,034 sequences) than the nr database, this search will run much faster than our previous search. Click on the BLAST button to begin your search.

8. How many hits came out of this search? Are all of them significant matches? (Note: in general, we consider matches with E-values less than 1 x 10^{-5} as statistically significant.) Does the BLAST result support the hypothesis that chimps have a homolog of the human *leptin* gene?

As you examine these alignments, you might notice that the accession numbers of the matches begin either with the NP_ or the XP_ prefix. A RefSeq prefix of NP_ indicates that the protein sequence record is supported by experimental evidence, while entries with an XP_ prefix are predicted computationally. Take a look specifically at the alignment to the match labeled "PREDICTED: pantothenate kinase 4 isoform X1 [*Pan troglodytes*]."

9. Explain why this alignment has a higher E-value than the alignment above it on the BLAST results list.

To further support your hypothesis that chimps have a homolog of the human leptin precursor gene, you can use the putative chimp leptin precursor sequence and do a blastp search against the human RefSeq protein sequence database. To do this, click on the GenPept link in the gray toolbar above the sequence you believe to be the chimp leptin precursor to retrieve the GenPept protein record. Click on the "FASTA" link under the gene name to obtain the full chimp protein sequence in FASTA format. Copy the sequence and go back to the NCBI BLAST page and click on the "Human" link under the "BLAST Genomes" section. Select the blastp program, verify that the "RefSeq protein" option is selected under the "Database" field, and then paste the chimp leptin precursor sequence into the "Enter Query Sequence" text box. Click on the BLAST button to perform your search.

10. Did you obtain matches from this search? Are they significant matches? How do you know? Does this data support your hypothesis about the presence of a leptin homolog in chimps?

1. Adapted from "BLAST Exercise: Detecting and Interpreting Genetic Homology" by W. Leung and SCR Elgin, Genomics Education Partnership, Washington University, St. Louis, MO.
2. Information accessed December 22, 2017.