Created by Paul Yenerall, 2010

**Using BLAST (Basic Local Alignment Search Tool) - blastp**

The purpose of this tutorial is to help teach students how to use and understand NCBI’s BLAST (Basic Local Alignment Search Tool). This tutorial will just cover blastp (blast protein) and will not go into depth about the use of BLAST in general. For more ways to use BLAST, please refer to the blastn tutorial as this tutorial goes more in-depth.

For questions, corrections, or help with this tutorial, please e-mail paulyenerall@gmail.comor lzhou1@pitt.edu

Before we begin, remember that blastp is used to search a protein sequence against a protein database. Any other type of query will require a different form of BLAST basic search. Refer to the blastn tutorial for more information on different types of BLAST queries.

**1)** First, load up an internet browser, preferably Mozilla Firefox, and go to the URL <http://blast.ncbi.nlm.nih.gov/Blast.cgi>. A page like the one listed below should appear:



Remember, we’re doing a protein sequence query, so select protein blast, as illustrated below:



The following page should be displayed after clicking “protein blast”. Note: most parts of this page will not be explained in great detail; if you would like a greater explanation on some of the features of this page, please refer to the blastn tutorial.



Now, please copy and paste this FASTA protein sequence into the query box labeled “Enter accession number, gi, or FASTA sequence”. Do not hit BLAST yet.

MNVEKAEFCNKSKQPGLARSQHNRWAGSKETCNDRRTPSTEKKVDLNADPLCERKEWNKQKLPCSENPRDTEDVPWITLNSSIQKVNEWFSRSDELLGSDDSHDGESESNAKVADVLDVLNEVDEYSGSSEKIDLLASDPHEALICKSERVHSKSVESNIEDKIFGKTYRKKASLPNLSHVTENLIIGAFVTEPQIIQERPLTNKLKRKRRPTSGLHPEDFIKKADLAVQKTPEMINQGTNQTEQNGQVMNITNSGHENKTKGDSIQNEKNPNPIESLEKESAFKTKAEPISSSISNMELELNIHNSKAPKKNRLRRKSSTRHIHALELVVSRNLSPPNCTELQIDSCSSSEEIKKKKYNQMPVRHSRNLQLMEGKEPATGAKKSNKPNEQTSKRHDSDTFPELKLTNAPGSFTKCSNTSELKEFVNPSLPREEKEEKLETVKVSNNAEDPKDLMLSGERVLQTERSVESSSISLVPGTDYGTQESISLLEVSTLGKAKTEPNKCVSQCAAFENPKGLIHGCSKDNRNDTEGFKYPLGHEVNHSRETSIEMEESELDAQYLQNTFKVSKRQSFAPFSNPGNAEEECATFSAHSGSLKKQSPKVTFECEQKEENQGKNESNIKPVQTVNITAGFPVVGQKDKPVDNAKCSIKGGSRFCLSSQFRGNETGLITPNKHGLLQNPYRIPPLFPIKSFVKTKCKKNLLEENFEEHSMSPEREMGNENIPSTVSTISRNNIRENVFKEASSSNINEVGSSTNEVGSSINEIGSSDENIQAELGRNRGPKLNAMLRLGVLQPEVYKQSLPGSNCKHPEIKKQEYEEVVQTVNTDFSPYLISDNLEQPMGSSHASQVCSETPDDLLDDGEIKEDTSFAENDIKESSAVFSKSVQKGELSRSPSPFTHTHLAQGYRRGAKKLESSEENLSSEDEELPCFQHLLFGKVNNIPSQSTRHSTVATECLSKNTEENLLSLKNSLNDCSNQVILAKASQEHHLSEETKCSASLFSSQCSELEDLTANTNTQDPFLIGSSKQMRHQSESQGVGLSDKELVSDDEERGTGLEENNQEEQSMDSNLGEAASGCESETSVSEDCSGLSSQSDILTTQQRDTMQHNLIKLQQEMAELEAVLEQHGSQPSNSYPSIISDSSALEDLRNPEQSTSEKAVLTSQKSSEYPISQNPEGLSADKFEVSADSSTSKNKEPGVERSSPSKCPSLDDRWYMHSCSGSLQNRNYPSQEELIKVVDVEEQQLEESGPHDLTETSYLPRQDLEGTPYLESGISLFSDDPESDPSEDRAPESARVGNIPSSTSALKVPQLKVAESAQSPAAAHTTDTAGYNAMEESVSREKPELTASTERVNKRMSMVVSGLTPEEFMLVYKFARKHHITLTNLITEETTHVVMKTDAEFVCERTLKYFLGIAGGKWVVSYFWVTQSIKERKMLNEHDFEVRGDVVNGRNHQGPKRARESQDRKIFRGLEICCYGPFTNMPTDQLEWMVQLCGASVVKELSSFTLGTGVHPIVVVQPDAWTEDNGFHAIGQMCEAPVVTREWVLDSVALYQCQELDTYLIPQIPHSHY

Now note that I did not provide any header; therefore, we do not know which protein this is, or from what species this sequence originates from. We will use blastp to find this out.

We do not need to change the database we are searching because unlike blastn, blastp always searches the non-redundant protein database, so it will search the best matches for all species. For this search however, let’s include a specific species we wish to search for. From the section of the page that says “Choose Search Set” go to the blank box that is labeled “Organism”, as illustrated below:



In this box, type in “Homo sapiens”, then choose the first option that should appear in the dropdown box (“Homo sapiens (taxid:9606)”), as illustrated below (next page):



Now, before clicking BLAST, click on the Algorithm Parameters link at the bottom of the page, which looks like this:


A page like the following page should be displayed. Pay particular attention to the highlighted areas as they will be discussed further:



1 – This part is the Word Size. The Word Size is how big of a match BLAST first finds to begin an alignment. For instance with a word size of 3, when BLAST finds three identical residue (amino acid) matches (between query and subject sequence) BLAST begins to see if it can match the rest of the query sequence to that location in the subject sequence. By turning this number down or up we can affect the sensitivity of matches

2 – This part is the Matrix. These substitution matrices assign a score for aligning any possible pairs of residues. The scores assigned for matches are based off of a large and diverse sample of verified pairwise alignments. The matrices are meant to reflect the probability of mutations occurring with evolution. The PAM (Point Accepted Mutations) matrix should be used for smaller queries because it is not very sensitive. The BLOSUM matrix (BLOcks SUbstituion Matrix) should be used for larger queries because it has greater sensitivity. For queries less than 35 residues, it’s best to use PAM-30. For queries between 35-50 residues, it’s best to use PAM-70. For queries between 50-85 residues, it’s best to use BLOSUM-80. For queries over 85 residues it’s best to use BLOSUM-62

3 – This part is the Gap Costs. This part assigns a penalty for each gap found in the alignment between the query and subject. By increasing the Gap Cost, we can find subjects that have less gaps, or vice-versa.

4 – This part is the Compositional Adjustments. The type of adjustment can be changed to compensate for the amino acids compositions of the sequences compared, which will change the score of the matches.

Note: You do not need to change any of these parameters for our query; however, you should be aware how and why you can change these parameters

Now click the BLAST button at the bottom of the page. Note that while searching a new box is displayed that was not displayed when doing a BLASTn search. The new box should look like the following picture:



This box will also be displayed on our BLAST results page. This box shows highly conserved regions between our query and possible matches.

Your results page should look something like this:


If you scroll down to the analytical results in the “Descriptions” part of the page we can get a better view of possible matches than the top Graphical output. As you can see, we have multiple very similar, very strong (query 100%, e-value = 0, high score) matches. However, these very strong matches are all linked to one gene, BRCA1. The reason there are multiple matches is because there are multiple isoforms of the BRCA1 gene in *Homo sapiens*.