How to BLAST

Performing a Search

1. Go to the NCBI BLAST page. This page lists all of the BLAST related tools/programs available from NCBI. For this exercise click on the link to the Standard protein-protein BLAST [blastp]. Please look over the page to become familiar with the information available and the data required for a search.

2. There are quite a few links on this page to information about BLAST and to explanations of the data required or the program settings. These links are a good resource and should be reviewed.

3. Paste the following FASTA format of the sequence into the large data entry field.

>unknown protein
MKNTLLKLGVCVSLLGITPFVSTISSVQAERTVEHKVIKNETGTISISQLNKNVW
VHTELGYFSGEAVPSNGLVLNTSKGLVLVDSSWDDKLTKELIEMVEKKFKKRVT
DVIITHAHADRIGGMKTLKERGIKAHSTALTAELAKKNGYEEPLGDLQSVTNLK
FGNMKVETFYPGKGHTEDNIVVWLPQYQILAGGCLVKSASSKDLGNVADAYV
NEWSTSIENVLKRYGNINLVVPGHGEVGDRGLLLHTLDLLK

4. The boxes immediately below the sequence entry box allow the selection of only a part of the entered sequence as the query for the search. For the purposes of this game, nothing should be entered in these fields.

5. The next field is a drop menu that allows the selection of the database to be searched. The non-redundant database (nr) is the default setting, but the new entries for the last month, EST's, a specific organism, or many others may also be selected. Leave the choice as nr for the game because it is the largest and most comprehensive database for BLAST to search.

6. The next check box allows the search of a conserved domains database. You may leave this checked, but the game will not use this function. Domain searches will be performed using BLOCKS and ProfileScan.

7. For basic searches, the only other option that should be changed is the filtering. Uncheck the box under the next section (Options) next to the words 'Choose filter.' For game purposes, the sequences should not normally be filtered. For information on situations when this box should be checked (only sometimes in levels 3 and 4), see the "More Advanced Information" section below.

8. Click on the "BLAST!" button to run the blastp search.

9. A new page will appear with the ID number of the search and the approximate wait time. Click on the "Format!" button and wait. The results will be returned when the search is complete.

Interpreting Results

1. The BLAST results page begins with the program version used, the reference for BLAST, the name and length of the query sequence, the database searched, and contact information. It is best to pay attention to this information so you will know that the proper query, database and program were run.

2. Next can be found the number of hits on the query sequence (i.e. the number of sequences in the database that have some similarity to the query sequence) and a graphical overview of the alignment of the hits to the query.

3. The long red line near the top of the graphical overview represents the length of the query sequence. Each hit from the database is represented by a line below the query that is colored to represent its score. The multicolored bar at the top of the graphical overview is a legend with the different colors representing different similarity score ranges. Red is used for hits with a score over 200 (good similarity) and so on.

4. The graphical overview shows the relative position of similar regions in each hit and the query and how big these regions of similarity are. Moving the mouse cursor over each hit line will bring up that hit sequences name in the text box above the graphic.

5. The graphical overview can be a very useful preliminary tool in the game to help determine if the sample (query) sequence is of terrestrial origin or not. However, do not make your determination only with this section of the results. If the query has hits that show up as long similar regions and are red and/or pink, it is probable that the query is terrestrial. If the query has hits that are short and black or blue, it is possible that the query is extra-terrestrial. If the query has short red or pink hits or hits that are in the middle of the range, more analysis than BLAST will need to be done to make an origin determination.

6. The next two sections of the results are more important, they show scoring and pairwise alignments. The section below the graphical overview ranks the hits by score. For each hit it gives a link by the accession number to the hit sequence in the database, the name of the hit sequence, the BLAST score, and the Expect value or E-value. The E-value is a statistical calculation based on the score that gives the number of hits of this score that this search would return by chance using a database of this size. I.e. If the E-value is 1, it is likely that you would get one chance hit with this score to the query using the particular database that was searched. The following general conclusions can be drawn from the E-values:

If the E-value is less than 1 X 10-50 ,the hit is very similar to the query sequence and is very likely to be evolutionarily related. If E-values in this range are obtained, the sequence can be assumed to be terrestrial contamination, especially if the names/descriptions of many of the top hits indicate that the hits are related to each other. However, at the higher game levels when more tools are available for use, the sample sequence should be analyzed with all tools to show corroborating evidence for the conclusion.

If the E-value is between 1 X 10-50 and 1 X 10-2, the hit has some similarity to the query sequence and may be related. When E-values in this range are obtained, the game sequence may be found to be terrestrial contamination, but further analysis will be needed. These values can indicate that the sample sequence is in the same family as the hit or it may have closely related functional domains. If the top hits all

seem to be related, this makes it more likely that the query is of the same family/type.

If the E-value is between 1 X 10-2 and 1, the hit has a slight possibility of being related to the query. This may indicate a distant evolutionary relationship. In order to conclude that the query sequence was either terrestrial or extra-terrestrial, much more analysis would be required.

If the E-value is above 1, the hit is not very closely related to any sequence in the database. If E-values in this range are obtained in the game, it can be concluded that the sample sequence is very possibly of extra-terrestrial origin. This conclusion can also be made when no matches are found at all.

7. For the search performed in this tutorial, the results give a top hit with an E-value of 1 X 10-132 which is an exact match to the query sequence. Sometimes exact and closely matched hits will have E-values of 0.

8. The last large section of the results, gives a pairwise alignment of the query sequence and similar regions in each hit. After the database accession number and name of the hit sequence, this section gives the score, E-value, percentage of bases or nucleotides that are identical and for proteins the percentage that are chemically similar (positives). The section is useful in determining if the query has an exact match in the database. If the percent identity of the first hit and the query sequence is 100% then the game sample sequence is in the database and the sample is definitely terrestrial contamination.

9. The very end of the result report gives information on the scoring matrix and gap penalties used, the number of sequences queried in the database and so on.

10. The query run for this tutorial results in a large number of full length and near full length red and pink hits shown in the graphical overview. The pairwise alignment with the top scoring hit shows a 100% identity with the query sequence. This leads to the conclusion that the query sequence is a Bacillus cereus beta-lactamase present in the nr database. If these results were obtained in the game, terrestrial contamination would be evident.

More Advanced Information

The information in this section is separated so as not to confuse the beginner, but will be useful and even necessary to the players in level 4.

Performing a Search

1. Go the the BLAST page again. This time choose the link to the Standard nucleotide-nucleotide BLAST and run a blastn search against the nr database with the following DNA sequence. This time, however, leave the 'Choose filter' box checked so that the sequence will be filtered for low complexity regions. Leave all of the other parameters at default as above. The filter option ensures that no false positive results are obtained due to short sequences that are very common across the spectrum of biological sequences. Be aware, however, that if the sequence is filtered, sometimes regions are X'ed out and an exact match sequence in the database may show only around 95% identity even though it is the same sequence.

>unknown DNA

CAGTCTAGTTCAAACTTACAAATCCTCAGAGTCCTCTTTTCGGCCATACACTTC
ACATCGGAAACATTAA ….

Interpreting Results

1. For the more complicated game problems presented in level 4, the length of BLAST hits becomes much more important.

If the query sequence is short (less than 100 nucleotides or amino acids long), the top E-values may be larger than 1 X 10-50 even if there is an exact match. Be sure and check the % identity of the top hits, not just the E-values.

Hits with low E-values that only have similarity to short regions of the query sequence are more likely to indicate that the sequences have motif or functional domain similarities rather than that they represent related genes or proteins. This is very likely the case when all of the matches are from sequences whose names and descriptions do not seem to indicate that the hits are in any way related to each other.

Hits with higher E-values, in the ranges of 1 X 10-50 to 1 X 10-5, may still indicate that the query and hits are related if the hit has at least a 35% identity with the query over at least 80% of its length. Another indication of this is if several hits have names and/or descriptions that indicate that they are related to each other. These should definitely be studied further.

2. Always look at the names/descriptions of the hits returned by BLAST. The higher the percentage of hits that seem to be related to each other, the more likely it is that the query is also related to them. But remember, if the similarity is only in small regions, this is more likely to indicate related functional domains rather than related proteins/gene products. If only a small region of a protein has some similarity to a few sequences in the database, it is a possibility that this resulted from convergent evolution (unrelated sequences evolving similar structures for similar functions) and the query might still be extra-terrestrial. This type of BLAST result would require much more study before any conclusion could be reached.

3. The blastn search performed on the sequence above results in some hits with mid-range E-values and the rest with higher ones. The top hits have over 80% identity to more that 80% of the length of the query sequence. And, all of the top matches are cat cytochrome b sequences, with many of the remaining matches coming from cytochrome b sequences of other mammalian species. Even though there is no exact match in the database, these results indicate that the sample sequence is almost surely terrestrial. Other available tools should be used to analyse the sequence to build a body of supporting evidence for this conclusion.

The five flavors of BLAST perform the following tasks:

blastp compares an amino acid query sequence against a protein sequence database;

blastn compares a nucleotide query sequence against a nucleotide sequence database;

blastx compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database;

tblastn compares a protein query sequence against a nucleotide sequence database dynamically translated in all six reading frames (both strands).

tblastx compares the six-frame translations of a nucleotide query sequence against the six-frame trans-lations of a nucleotide sequence database.